



Article

Smart Building Energy Inefficiencies Detection through Time Series Analysis and Unsupervised Machine Learning

Hanaa Talei ^{1,2,*}, Driss Benhaddou ³, Carlos Gamarra ⁴ , Houda Benbrahim ⁵ and Mohamed Essaïdi ⁶ ¹ Smart Systems Lab, ENSIAS, Mohammed V University in Rabat, Rabat 10000, Morocco² School of Science and Engineering, Al Akhawayn University, Ifrane 53000, Morocco³ Department of Computer Engineering Technology, University of Houston, Houston, TX 77204, USA; dbenhadd@central.uh.edu⁴ Houston Advanced Research Center, Houston, TX 77381, USA; cgamarra@harcresearch.org⁵ Department of Computer Science and Decision Support, Ecole Nationale Supérieure d'Informatique et d'Analyse des Systèmes, Rabat 10112, Morocco; benbrahimh@hotmail.com⁶ College of Engineering, Ecole Nationale Supérieure d'Informatique et d'Analyse des Systèmes, Rabat 10112, Morocco; essaïdi@gmail.com

* Correspondence: h.talei@aui.ma

Abstract: The climate of Houston, classified as a humid subtropical climate with tropical influences, makes the heating, ventilation, and air conditioning (HVAC) systems the largest electricity consumers in buildings. HVAC systems in commercial buildings are usually operated by a centralized control system and/or an energy management system based on a fixed schedule and scheduled control of a zone setpoint, which is not appropriate for many buildings with changing occupancy rates. Lately, as part of energy efficiency analysis, attention has focused on collecting and analyzing smart meters and building-related data, as well as applying supervised learning techniques, to propose new strategies to operate HVAC systems and reduce energy consumption. On the other hand, unsupervised learning techniques have been used to study the consumption information and profile characterization of different buildings after cluster analysis is performed. This paper adopts a different approach by revealing the power of unsupervised learning to cluster data and unveiling hidden patterns. In this study, we also identify energy inefficiencies after exploring the cluster results of a single building's HVAC consumption data and building usage data as part of the energy efficiency analysis. Time series analysis and the K-means clustering algorithm are successfully applied to identify new energy-saving opportunities in a highly efficient office building located in the Houston area (TX, USA). The paper uses 1-year data from a highly efficient Leadership in Energy and Environment Design (LEED)-, Energy Star-, and Net Zero-certified building, showing a potential energy savings of 6% using the K-means algorithm. The results show that clustering is instrumental in helping building managers identify potential additional energy savings.

Keywords: smart building; LEED building; energy efficiency; unsupervised learning; clustering; time series; IoT



Citation: Talei, H.; Benhaddou, D.; Gamarra, C.; Benbrahim, H.; Essaïdi, M. Smart Building Energy Inefficiencies Detection through Time Series Analysis and Unsupervised Machine Learning. *Energies* **2021**, *14*, 6042. <https://doi.org/10.3390/en14196042>

Academic Editor: Robert Černý

Received: 14 July 2021

Accepted: 20 August 2021

Published: 23 September 2021

Corrected: 27 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

According to the US Department of Energy, commercial buildings consume approximately 20% of the United States' energy [1]. In addition, it is well documented that HVAC consumes more than 50% of the overall energy in commercial buildings [2]. The potential cost savings and desirable environmental impacts have enticed many researchers from various fields to investigate ways to optimize energy consumption. Given the significance of the amount of power consumed by HVAC systems, several technological developments have aimed to operate HVAC systems more efficiently. However, there is still room for improvement in the management and operation of HVAC systems, even in the most highly efficient certified buildings [3]. This improvement will be possible by

considering weather data, building-related data and machine learning algorithms, and following data-driven approaches.

Internet of Things (IoT) devices (sensors, meters, actuators, etc.) make it possible to access data related to building attributes. From a building management perspective, the IoT and Artificial Intelligence (AI) provide more saving opportunities through machine learning-based smart energy management techniques by considering how the occupants use the building [4]. Significant attention has been paid to collecting and analyzing occupancy data, as researchers strongly believe that adjusting HVAC operations to consider the occupants' presence and patterns will lead to significant energy savings [5–7]. However, research that uses supervised machine learning (ML) techniques to adjust HVAC operation to occupant activities assumes that the user has direct control of the HVAC system, making it applicable only to homes or small offices [7]. In addition, the supervised ML techniques tested for commercial buildings' HVAC systems are mainly used for the prediction of consumption or baselining to understand and evaluate the impact of the energy-saving adjustment implemented by the building managers [8]. This paper adopts a different approach, as it demonstrates that an unsupervised ML technique offers the opportunity to identify and measure inefficiencies in the way buildings are managed. Building managers can use these findings to develop more energy-efficient strategies and adjust the Building Energy Management system (BEMS), a computer-based system used by energy managers to control and optimize the energy needs of a system, accordingly [9].

Previous publications have utilized unsupervised machine learning techniques to identify typical operation patterns in buildings [10–18]. Most of these papers have focused on the reliability and the sensitivity of knowledge discovery and how it can be used to operate buildings better. Fan et al. used an Entropy-Weighted K-means (EWKM) clustering algorithm to identify consumption patterns and used the data to cluster weekdays and weekend operations [9]. Other data mining techniques have been utilized to discover hidden knowledge in a large dataset to improve building operational performance [19,20]. However, domain knowledge expertise is still needed to apply the knowledge and discover the possible performance improvement of the building operation. Among these papers, Howard et al.'s work is the closest to what we propose in this paper. Howard et al. developed an automated data mining method for identifying energy efficiency opportunities using whole-building electricity data. The researchers utilized a two-step approach, using piecewise linear regression and the density-based robust regression model residual clustering to detect both schedule- and operation-related electricity consumption faults [14]. The study is limited to identifying inefficiencies caused by operation- and schedule-related faults, representing an obvious deviation from what is expected to be considered a regular operation. Our paper is different from the previous work in that it uses unsupervised ML not only to understand the energy consumption patterns of a building, but also to identify energy inefficiencies in an HVAC system and quantify potential energy savings, even in a highly efficient building. Our work demonstrates the importance of using human activity in managing HVAC systems and shows the associated savings. In addition, the technique shows the inefficiencies without the need for subject matter expertise.

This paper uses 1-year timeseries data of a Platinum LEED ENERGY STAR 99/100, Net Zero-certified building in the Houston, TX area [21]. From the building design, construction, and management perspectives, the LEED standards were developed in 1993 to build highly efficient and green buildings that are comfortable and healthy for occupants. From an operational standpoint, ENERGY STAR certification is a public benchmarking system that helps commercial energy users understand how their building is ranked among similar buildings in the United States [3]. LEED buildings use BEMS to manage HVAC systems and adjust the building's thermal comfort based on occupant preference. This preference is configured by the building manager following the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) standards [22].

The main contributions of the paper are:

1. This paper presents, in detail, a data analysis process applied to data collected from Houston Advanced Research Center (HARC).
2. This paper demonstrates the significance of using unsupervised ML techniques to identify inefficiencies.
3. This paper demonstrates the importance of using occupant activities in the way building is managed.
4. This paper quantifies the possible energy savings that AI can provide in the operations of HVAC systems, even in highly efficient buildings.

The rest of this paper is structured as follows. Section 2 describes the thermodynamic process used by an HVAC system and presents a literature review of the most common supervised learning techniques used to decrease HVAC energy consumption. Data collection, description, and exploration are presented in detail in Section 3. Section 4 discusses how the K-means clustering algorithm works, different metrics used to pick the right number of clusters, and the importance of data scaling. The results of applying the K-means algorithm to the data analyzed in this paper are outlined in Section 5. Finally, the conclusions and future work are presented in Section 6.

2. Heating, Ventilation, and Air Conditioning (HVAC): An Overview and Literature Review

Nowadays, many HVAC systems are developed to match building types, mercantile and services, offices, health care, education, and lodging requirements. Other key elements that impact the choice in HVAC system include the building architecture, climate conditions, energy efficiency objectives, and the owner's preference. In a typical LEED building, the choice of a highly efficient HVAC system is significant as it affects its overall LEED scoring. In gold- and platinum-certified facilities, renewable energy technologies such as solar and geothermal are utilized. In particular, geothermal technology uses groundwater heat pumps for direct radiant cooling and heating and has one of the highest Seasonal Energy Efficiency Ratio (SEER) certifications granted to efficient central air conditioning systems owners [23].

The data-driven approach requires the acquisition of data at different operational points of an HVAC system. In addition, energy-efficient operations require the adjustment of some HVAC operating points. Therefore, it is essential to understand how an HVAC system works, its components, and the different steps of its thermodynamic cycle, as presented in Figure 1.

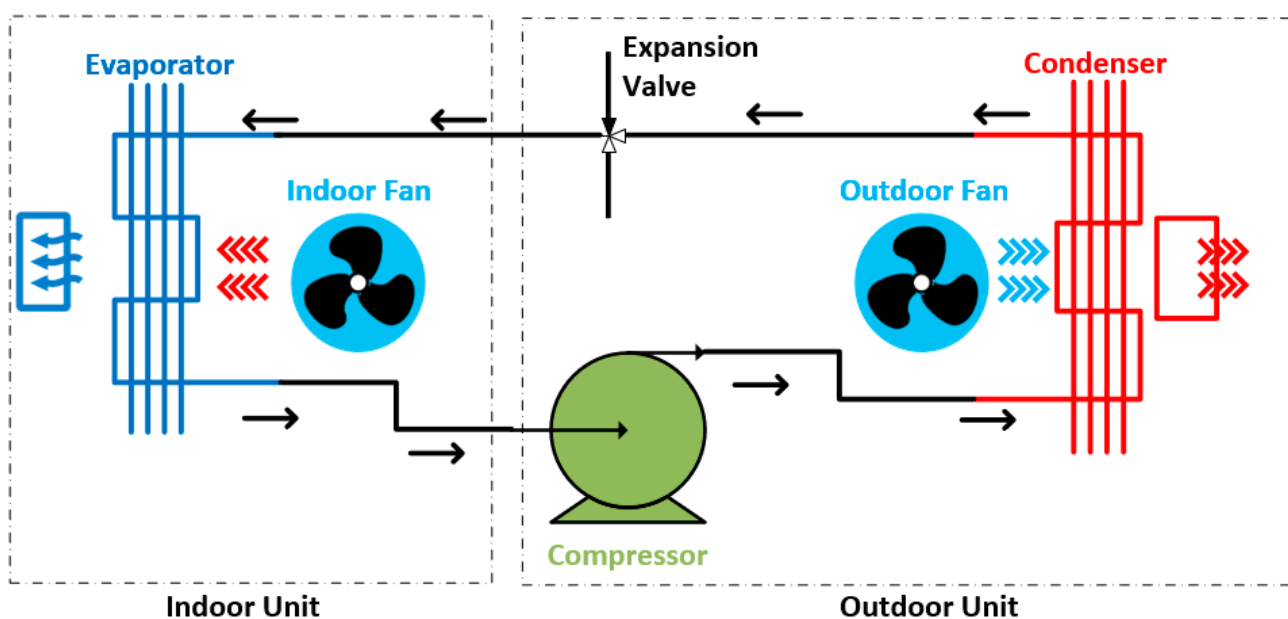


Figure 1. HVAC thermodynamic process.

An HVAC system consists of four major components: a compressor, a condenser, an expansion valve, and an evaporator. HVAC systems follow a thermodynamic process to remove heat from one place, replace it with cold air, and expel the hot air to the outside atmosphere. To achieve this exchange of heat, HVAC activity requires the use of a fluid called refrigerant that flows through the connecting pipelines and the parts of an HVAC system. Refrigerant changes liquid to vapor and vice versa at convenient temperatures during the HVAC thermodynamic process. In physics, the second law of thermodynamics states that heat flows from an object at a higher temperature to an object at a lower temperature. To achieve this, a compressor is used to compress the refrigerant, thus increasing its pressure/temperature compared to the outdoor pressure/temperature. The high pressure/temperature gas refrigerant continues its journey and reaches the condenser, where it changes its physical states from gas to liquid as the heat energy is absorbed from the hot gas. Then, the heat energy is expelled to the outside environment using the outdoor fan. The high temperature condensed liquid refrigerant reaches the expansion valve, which reduces its molecules' pressure, cools down its temperature, and controls the amount of the refrigerant that enters the evaporator. The latter is a cooling effect generator, as the cold refrigerant absorbs the heat from the indoor temperature and starts to evaporate to form a vapor. Then, the indoor fan circulates the cold air from the coiling surface to the indoor environment. Finally, the vapor leaves the evaporator and moves to the compressor where the cycle is repeated [24].

Generally, building managers can optimize HVAC operations by adjusting the thermostat temperature as a function of the number of people in a building and occupied space, which drives the amount of air to be pumped in the space, therefore influencing the power consumption. Most building managers set up a schedule based on the business operation times: day vs. night and weekdays vs. weekend. However, HVAC systems can still exhibit inefficiencies, as some space in the building is not occupied during business hours, and the number of people fluctuates during the operation of the building. Ongoing ML-based research is oriented toward identifying HVAC inefficiencies, using supervised learning techniques to propose strategies to operate HVAC systems to decrease energy consumption while maintaining adequate indoor comfort and then measuring the performance of the proposed strategy. Using a model predictive control algorithm and adopting a logical regression model to predict and adjust setpoints, researchers have proposed different strategies to improve the use of HVAC systems.

Model predictive control (MPC) is a feedback control algorithm that uses a model to make predictions about future outputs of a process by handling multiple input data and constraints. Given data availability, the research community is interested in applying MPC for efficient energy management in buildings. The problem formulation, applications, and opportunities of various MPC algorithms managing an HVAC system have been presented by G. Serale [25]. In their study, the researchers addressed the problems of energy waste and inefficient operation of the HVAC system, which occur due to the lack of a reliable system that can measure and predict the building occupancy. To solve this problem, the researchers developed a logistic regression model to forecast the building occupancy. The forecasted occupancy was used among the key variables to consider in the MPC framework. Numerical simulations were applied to a building using 73-day, 15-minute interval occupancy data collected from a low-income residential house in San Antonio, Texas. The results proved that a potential energy savings of 8% could be achieved while maintaining occupant comfort. The work presented fits in developing supervised learning techniques as a strategy to operate HVAC systems. However, the work does not describe how the HVAC inefficiencies were identified.

Besides the use of MPC, W. Li addressed the energy improvement of an HVAC system using an optimal setpoint and turning off the HVAC system automatically in unoccupied offices [26]. The researchers deployed a set of IoT-based sensors to detect the indoor environmental quality in real time, in which the temperature and CO₂ concentration were considered. The collected temperature data were used to identify if the HVAC setpoint

was maintained within a preferred range, while the CO₂ concentration data were used to check if the HVAC system was turned off during unoccupied timeslots (i.e., low CO₂ concentration). To estimate an optimal setpoint temperature of the HVAC system, the authors used the mean value of the indoor temperature, utilizing one of the change point analysis (CPA) models to evaluate the actual status of the setpoint temperature used in the HVAC system. On the other hand, the indoor environmental indicators were also used to evaluate whether the setpoint temperature was effective in turning off the HVAC system in an unoccupied space. The conducted analysis tackles some important factors that impact the HVAC system efficiency. However, other important factors, such as outdoor temperature and actual HVAC consumption, were not considered in the analysis to find the inefficiencies.

The impact of using an optimal setpoint on energy conservation was also investigated by S. Papadopoulos [27], who adopted an optimization framework to finetune heating and cooling setpoints of large office buildings with respect to energy consumption and occupant thermal comfort. The researchers simulated the building performance using a multi-objective optimization algorithm. The study used genetic algorithms to evaluate the objective functions and provide solutions for the parameters of each problem. These parameters were then used to define the optimum setpoints. The performance of these methods was tested in seven different climate zones in the US and showed a 60% annual HVAC related energy savings without affecting the occupants' thermal comfort. Despite promising objectives of adopting the proposed framework, the work used only a simulated environment and did not use real data to analyze the behavior of the test building. The work also lacks a systematic approach to find the inefficiencies.

In contrast to the previous work, H. Do [28] conducted data-driven research evaluate a building's HVAC efficiency using building assessors, electricity demand, and outdoor environmental data. The building assessors' data were used to estimate the HVAC system size and then estimate the HVAC system's electricity demand curve. To define if the HVAC system was operating as expected, the authors proposed an HVAC efficiency rating to compare the model's predictions and the actual performance data and considered as a case study 39 occupied residential buildings in Austin, Texas. The study concluded that 85% of the analyzed buildings had an efficient HVAC system, while 15% did not. The proposed HVAC efficiency rating can help to identify HVAC systems in need of energy efficiency upgrades.

W. Jung [29] addressed the importance of reducing unnecessary energy consumption by considering the human-in-the-loop in HVAC operations. The authors used a structured literature review approach, which involved investigating the human-in-the-loop according to two human dynamics parameters that drive user-centric operations of HVAC systems: occupancy and comfort. The paper proposed a five-tier hierarchical taxonomy (human-in-the-loop HVAC modality, building type, measurement techniques, sensing performance, and HVAC performance). The studies in the paper were classified based on their contributions to occupancy- and comfort-driven human-in-the-loop HVAC operations. Besides, the authors distinguished simulations from field evaluations to assess the actual viability and challenges in achieving the desirable results in practice. The authors also used a hype cycle model to qualitatively evaluate the developments of different technologies for human-in-the-loop HVAC operations from a research perspective. The identification of energy-efficient opportunities relies mainly on the human-in-the-loop experience.

Many researchers strongly believe that accounting for the dynamics of occupancy in HVAC operations will lead to important energy savings and improve indoor thermal comfort. To achieve this, D. Ardiyanto [5] proposed a method to adjust HVAC setpoints based on occupant comfort, which is measured after computing the hourly Predicted Mean Vote (PMV) based on real-time occupancy information, indoor temperature setpoints, and humidity in a building. To test the effectiveness of the proposed method, the authors developed a building model based on a real one located in Alexandria and simulated the electrical consumption behavior after increasing the setpoint during unoccupied slots.

The proposed method proved that more than 23% electrical consumption could be saved. However, V. Erickson [30] proved that it is possible to achieve an annual energy savings of 42% if real-time occupancy data are used.

As noted, most existing of the work on energy efficiency has focused on using supervised machine learning techniques to implement HVAC operating strategies. However, these studies have not considered how to systematically determine the inefficiencies. Even though expert experiences are valuable and show efficiency improvement, they may miss opportunities to achieve more efficiency. Methods must further explore the inefficiencies missed by experts. This paper demonstrates the importance of collecting and analyzing smart meter data, as well as how unsupervised machine learning can be utilized to find inefficiencies in the way an HVAC system is operated.

3. Data Collection, Description and Exploration

3.1. Data Collection and Description

The unsupervised ML technique adopted in this paper used the time series power consumption data collected from the Houston Advanced Research Center building, an 18,600 square foot office building located in The Woodlands, Texas (USA), as pictured in Figure 2. Designed and certified as a LEED Platinum building, the HARC operates the building as a living lab and continues pushing its boundaries. The HARC is one of the 55 office buildings in the US certified as Zero Energy and has an Energy Star score of 100 out of 100.



(a)



(b)

Figure 2. The Houston Advanced Research Center Building: (a) outside view, (b) photovoltaics array.

The power consumption data were collected at different building levels and included the whole building and four sub-systems: lighting, plug loads, HVAC, and others. We followed the standard data analytics process, presented in Figure 3, which consists of defining the problem to solve, collecting the data, exploring the data, proposing a solution to the target problem, and evaluating the proposed solution. This analysis aims to study the building energy consumption and determine whether it is possible to identify any inefficiencies in the HVAC system.



Figure 3. Data analysis steps.

The data analyzed in this paper consisted of 1-year data collected from four meters that capture 1 min consumption of an HVAC system, lighting, plug loads, and others. Table 1 describes the raw data and wrangled data used in the analysis.

Table 1. HARC collected data—a description.

Data Aspect	Description
Before Wrangling	
Initial data Dimension	12 by 525477
Description	1 min interval data that represent the following features: date, timeslot, others in kilowatt (kW), plug loads in kilowatt (kW), lighting in kilowatt (kW), HVAC consumption in kilowatt (kW), total in kilowatt (kW), others in kilowatt hour (KWh), plug kilowatt hour (KWh), lighting in kilowatt hour (KWh), HVAC consumption in in kilowatt hour (KWh), total in kilowatt Hour (KWh)
Sampling Rate	1 min
Number of rows missing	123
After Wrangling	
Final Dimensions	3 by 8760
Sampling Rate	1 h
Number of rows	8760
Data Span Period	From October 2018 to September 2019

3.2. Data Preparation and Exploration

Before exploring the data, we followed steps to prepare it:

1. We replaced a missing row with the average of the previous two rows.
2. If more than one row was missing, and because the consumption was a pattern repeated weekly, we replaced the missing rows a similar day's timeslot data.
3. We changed the data sampling rate from a 1 min interval to a 1 h interval for better visualization of the data.

Descriptive statistics and plot exploratory graphs are essential to understand data. Figure 4 presents the energy consumption of four meters both for November 2018 (a cold month in Houston) and August 2019 (a hot/humid month in Houston). As expected, and because Houston weather is classified as a humid subtropical climate, the HVAC meter showed a high consumption percentage compared to the other meters values. We decided to narrow our analysis to study how the HVAC system is operated in the Houston Advanced Research Center.

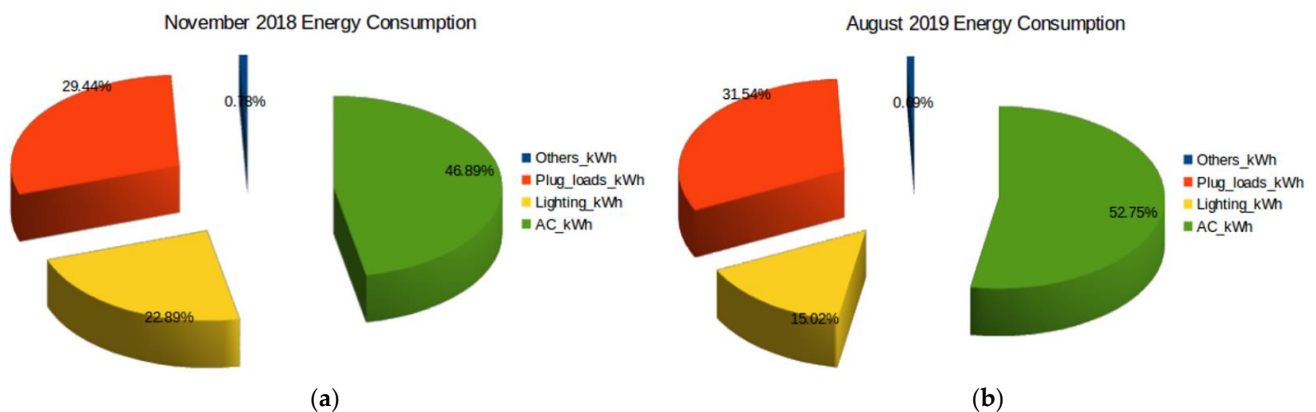


Figure 4. HVAC consumption: (a) November 2018, (b) August 2019.

The Houston Advanced Research Center uses a centralized HVAC system called Webcontrol, developed by the company Automated Logic. The HVAC system is controlled by changing a setpoint following a scheduled occupancy of the building and depending on the climate period (i.e., hot (April to October) or cold (November to March)). To evaluate the efficiency of the HVAC system, it was important to consider other essential variables in the analysis, such as the outdoor temperature, indoor temperature, setpoint schedule, and number of occupants. Unfortunately, all these variables' values were not recorded. However, given the importance of these data in the analysis, we collected the outdoor temperature data from the WILLIAM P. HOBBY AIRPORT STATION weather station, which provides hourly temperature data. We were not able to use the indoor temperature in the analysis. As for the number of occupants, given that access to the building involves the use of access cards, the system records only the entry time for every user. We used these data to count the number of occupants in the building and assumed that every user would remain in the building until 17:30 h. We did not use the setpoint in the analysis, as the value was changed by the building energy managers when needed.

After adding data on the outdoor temperature and the number of users, we plotted a first graph of the 1-month data for October 2018. The graph in Figure 5 represents the daily values of the HVAC power consumption, the number of users, and the outdoor temperature scaled in the range of 0 and 1. We identified that on 8 October, the HVAC consumption was high while the number of users was relatively small, which might be explained by a change in the setpoint or an inefficiency.

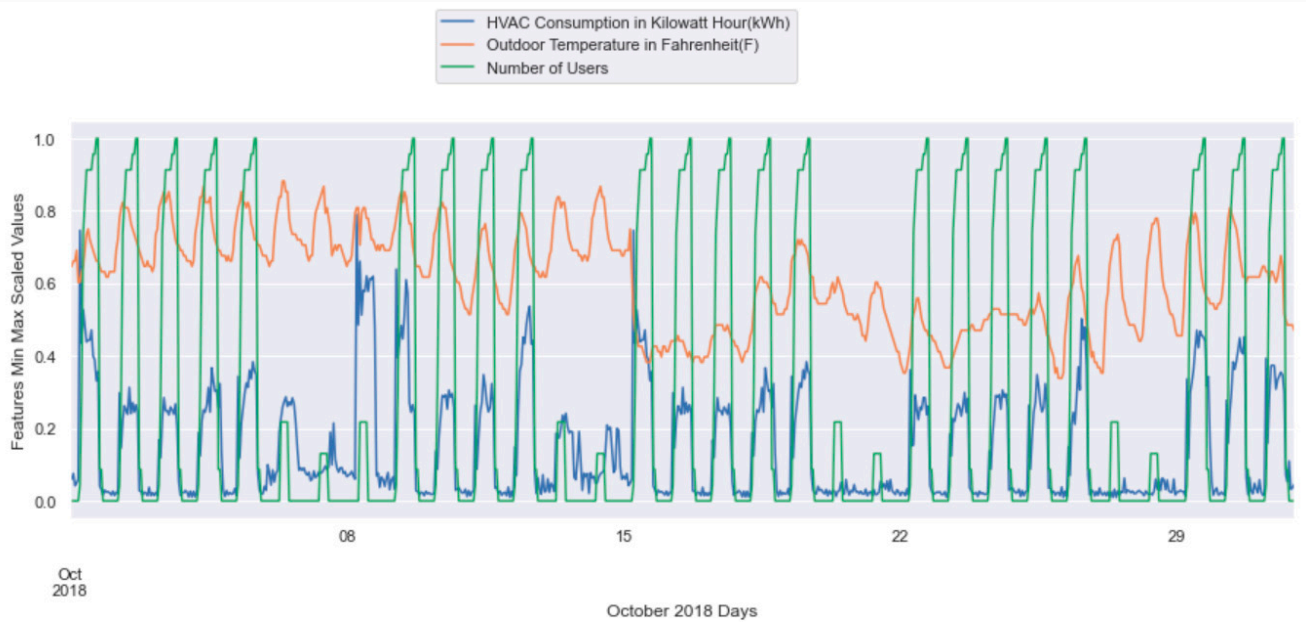


Figure 5. October 2018 data representation.

To understand the change of HVAC consumption over the week, we ran the autocorrelation function (ACF) on the hourly representation of HVAC power consumption. ACF determines the autocorrelation between a time series and the same time series offset by n steps, which can be plotted to obtain an overview of the data [31]. The autocorrelation used the Pearson coefficient (r), which was calculated as follow [32]:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}} \quad (1)$$

where n is number of datapoints, x is the value of the x -variable in a sample, and y is the value of the y -variable in a sample.

The value of the autocorrelation coefficient ranges between -1 and 1 , where the latter means a perfect positive correlation and the former means a perfect negative correlation. Figure 6 is the ACF plot of the HVAC consumption, where the bars show the ACF values at increasing timesteps (lags). Excluding the first bar that compares the autocorrelation of variable value with itself (value obtained: 1), the highest correlation was obtained at lag (timestep) 168, which implies that the HVAC consumption has a weekly repeated behavior.

To explore the change of HVAC consumption throughout the day, we plotted the box graph of hourly consumption of the whole-year data in Figure 7. Because the setpoint was changed at 6:00 a.m., Figure 7 clearly shows a change of consumption starting at 6:00 a.m., which then stabilizes over the next 3 h. Beginning at 10 a.m., the temperature in Houston starts to increase, reaching a peak at 3:00 p.m. Then, the temperature decreased afterward, stabilizing around 8:00 p.m. These patterns are clearly represented in the figure. It is noteworthy that the HVAC consumption dropped after 16:00 and then continued to decrease after 17:00 because the setpoint was changed again.

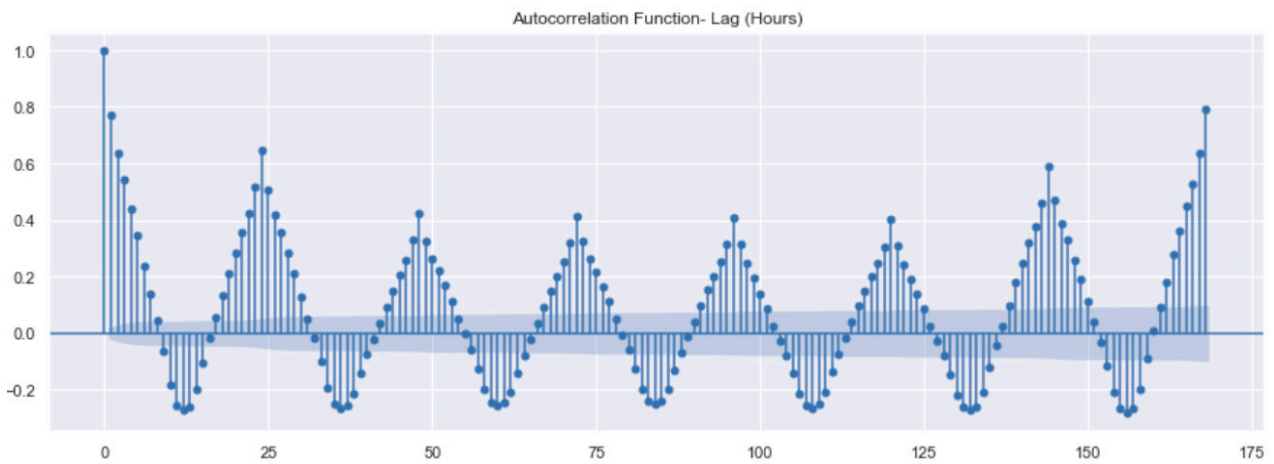


Figure 6. HVAC consumption-ACF plot.

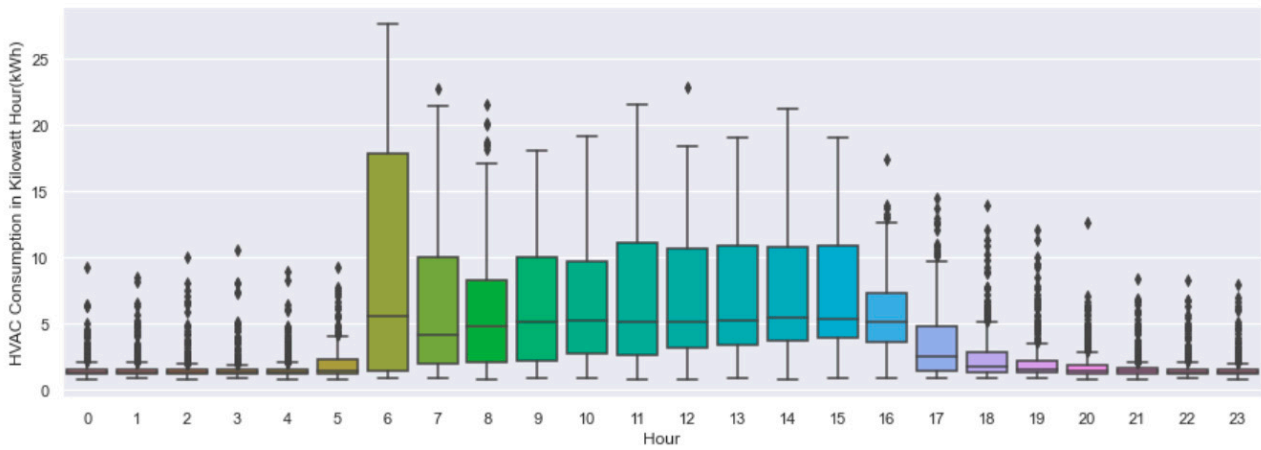


Figure 7. HVAC consumption change per hour.

After exploring the change of HVAC consumption throughout every hour/day, we looked at the consumption change by weekday. As expected, and given the setpoint configuration, the HVAC consumption exhibited the same behavior during the working days (with a slight decrease on Fridays). The same trend applied for weekend days, as presented in Figure 8.

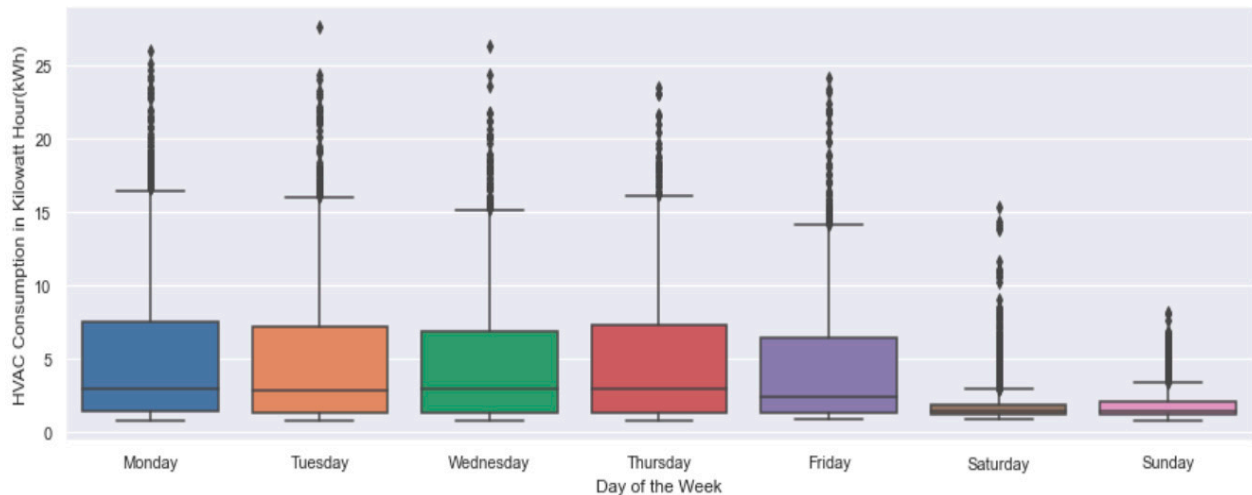


Figure 8. HVAC consumption per days of the week.

We also looked at the change of HVAC energy consumption by considering the number of the users and the outdoor temperature. Table 2 presents the distribution of the number of users throughout the day for working, weekend, and vacation days.

Table 2. Users distribution per hours of the day.

Day/Time Slot	00:00	06:00	07:00	08:00	09:00	10:00	14:00	16:00	18:00	20:00
	05:59	06:59	07:59	08:59	09:59	13:59	15:59	17:59	19:59	23:59
Working	0	1	7	17	19	21	22	23	2	0
	From 00:00 to 7:59			From 8:00 to 12:59			From 13:00 to 23:59			
Saturday	0			5			0			
Sunday	0			3			0			
Vacation	0			5			0			

To explore the change of HVAC consumption by considering both the number of users and the outdoor temperature, we plotted a matrix representing the correlation between the three features, as presented in Figure 9.

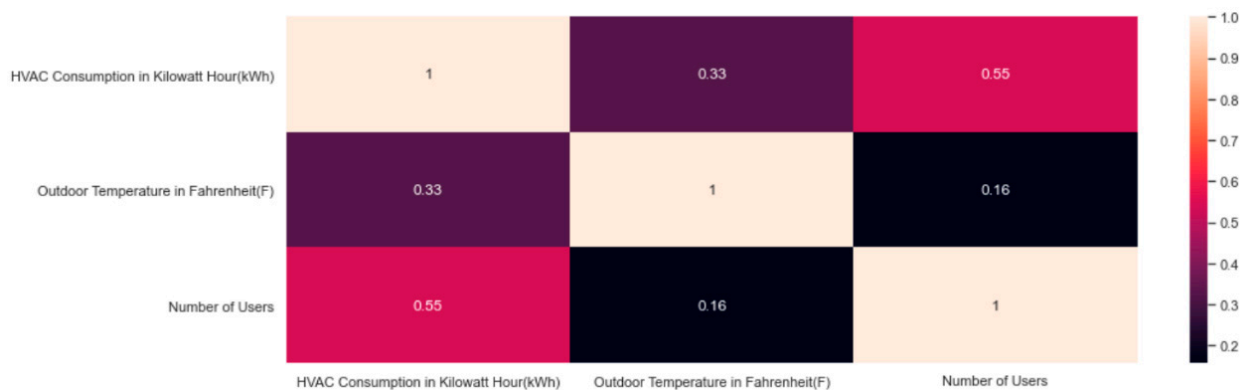


Figure 9. Correlation coefficients between the HVAC consumption, number of users, and outdoor temperature.

Unexpectedly, the correlation between the HVAC consumption and the outdoor temperature was not very high, which suggests that other factors affected the power consumption. The exploration phase unveiled a high consumption when the number of users or the outdoor temperature was low. Therefore, we decided to use the K-means algorithm to identify the timeslots that explain the HVAC consumption behavior.

4. K-Means Clustering Algorithm: A Step-by-Step Process

Machine learning algorithms can be split into two main categories: (1) supervised learning, which involves making predictions with the help of labeled datasets (for instance, given the geometrical measurement of a house and its location, a supervised algorithm could predict its price); and (2) unsupervised learning, which involves using unlabeled data to extract some information/patterns (for instance, finding out the probability of the co-occurrence of items in a collection [33]). Besides associations, unsupervised learning is also involves solving clustering problems meant to divide data into groups, where every group contains data with the same behavior. For instance, after collecting consumption data using building meters, an energy manager is eager to find buildings with the same consumption behavior.

To identify inefficiencies in the HVAC system for the Houston Advanced Research Center, the K-means algorithm was used to cluster the time series data presented in the previous section. K-means is a clustering algorithm used to group data into K groups using the mean (average) computation [34]. It is implemented in many data analysis tools, and in this paper, we used the version implemented in the Scikit-Learn Python package. Figure 10

is a flowchart that presents the steps used in the K-means algorithm to group data into k clusters.

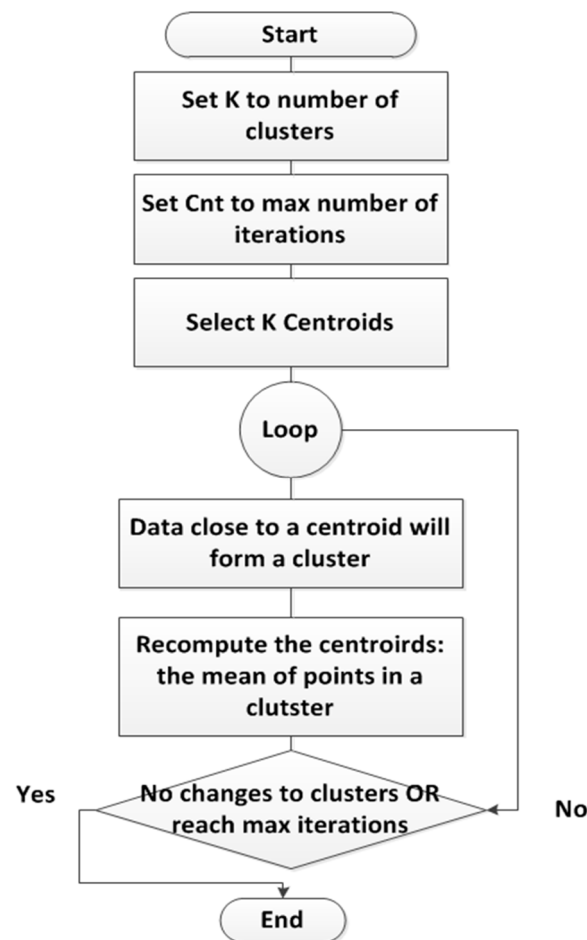


Figure 10. Main steps of the K-means algorithm.

The simplicity of this algorithm is among its “raison d’être” and why it is widely used. K-means generally produces satisfactory clustering results, yet it has some caveats that are addressed in this paper. The first caveat is the random initialization of the centroids, which can sometimes lead to locally optimal solutions that are not globally supported. It is important to mention that different initializations generate different clustering results, which are sometimes far from the optimal. The second caveat is the determination of the number of clusters before running the algorithm [35].

To avoid the first caveat, we used the Scikit-Learn K-means implementation, which uses a method of initialization based on k-means++ algorithm that chooses the centroids in a smart way to speed up the convergence [36]. K-means ++ consists of the following steps:

1. Randomly choose a datapoint to be a cluster center
2. For every other datapoint, compute the distance (let’s call it $D(x)$) from a point to the cluster center
3. Choose the next centroid such that the probability of choosing a point is proportional to $D(x)^2$
4. Repeat the above two steps until the right number of centroids are found

Regarding the second caveat, the algorithm classifies points according to the specified number of clusters k . To choose the right k value, different metrics are used to assess the clustering results using different numbers of clusters. The ideal results occur when the inter-cluster is minimized and the intra-cluster is maximized [34]. For all the indices used in the analysis, the optimal number of clusters was based on the location of a bend in the

generated graphs. In this paper, we used four different techniques that are explained in Table 3.

Table 3. Different clustering metric techniques used in the analysis.

Metric Name	Description	Mathematical Formula	Interpretation of Score
Elbow Method [37]	Looks at the within-cluster sum square (measures the intra-cluster variation) as a function of the number of clusters	$WSSE = \sum_{j=1}^k \sum_i^n \ \mu_j - x_i\ _2^2$	Lower is better (homogenous clusters), as a higher score represents more heterogeneous clusters
Silhouette Score [38]	Used to study the separation distance between the resulting clusters	Silhouette Score = $\frac{x-y}{\max(x,y)}$ x : mean distance to the points in the closest cluster y : the mean intra cluster distance	Silhouette Score value is in the range of (1, 1) +1 indicates that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters and negative values indicate that those samples might have been assigned to the wrong cluster.
Davies Bouldin Score [39]	The average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances.	Davies Bouldin Score = $\frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \frac{icd(C_i) + icd(C_j)}{\text{distance}(C_i, C_j)}$ where icd refers to intra cluster distance	Minimum score is 0 with lower values indicating better clustering
Calinski Harabasz Score [40]	The score is defined as ratio of the sum of between clusters dispersion and of inter-cluster dispersion	Calinski Harabasz Score = $\frac{n-k}{k-1} \cdot \frac{bcd}{icd}$ where bcd refer to between cluster dispersion and icd refer inter cluster dispersion	Higher score refers to better defined clusters

In addition to choosing the right number of clusters, feature scaling is an important step for many machine learning algorithms that use the distance between data [41]. Commonly, the K-means algorithm compares data using the Euclidean distance using the following formula:

$$\text{Euclidean Distance } (p1(x_1, y_1), p2(x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (2)$$

where $p1$ and $p2$ are the two points in Euclidean Space; and x, y are the Euclidean vectors starting from the origin of the space.

Since data clustering is based on the difference between the value points, it is important to have all features on the same scale. Otherwise, high values will be grouped regardless of their patterns. To remove the volume difference in the data, we used three different scaling methods and chose the method that led to balanced clusters. Table 4 describe the scaling methods used from the Scikit-Learn package.

Table 4. Data scaling methods used in this paper.

Scaling Method	Description	Mathematical Formula
Min-Max Scaler	Used to normalize data in the range of [0,1] For each value in the feature, the minimum value is subtracted and then divided by difference between the original maximum and original minimum [42]	$\frac{X - \min}{\max - \min}$
Standard Scaler	Used to rescale the distribution of the data by subtracting the mean and then dividing by the standard deviation [43]	$\frac{X - \mu}{\sigma}$
Robust Scaler	Primary used to remove the effect of outliers as the centering and scaling of this scaler are based on percentiles [44]	$\frac{x - Q_1}{Q_3 - Q_1}$

Given that three different features were used in our analysis, different units were used, and the range of raw data values varied widely, it was important to standardize the values in a specific range. The effect of scaling data is clearly shown in Figure 11, which presents a sample of the original data (18 June 2019 data) along with the impact of the scaling methods used. The latter clearly removed the difference in volumes and revealed some initial data patterns, such as the mutual change of HVAC consumption, along with

the change of the number of users and temperature, excluding the timeslots where the setpoint was changed (i.e., 06:00 a.m.).

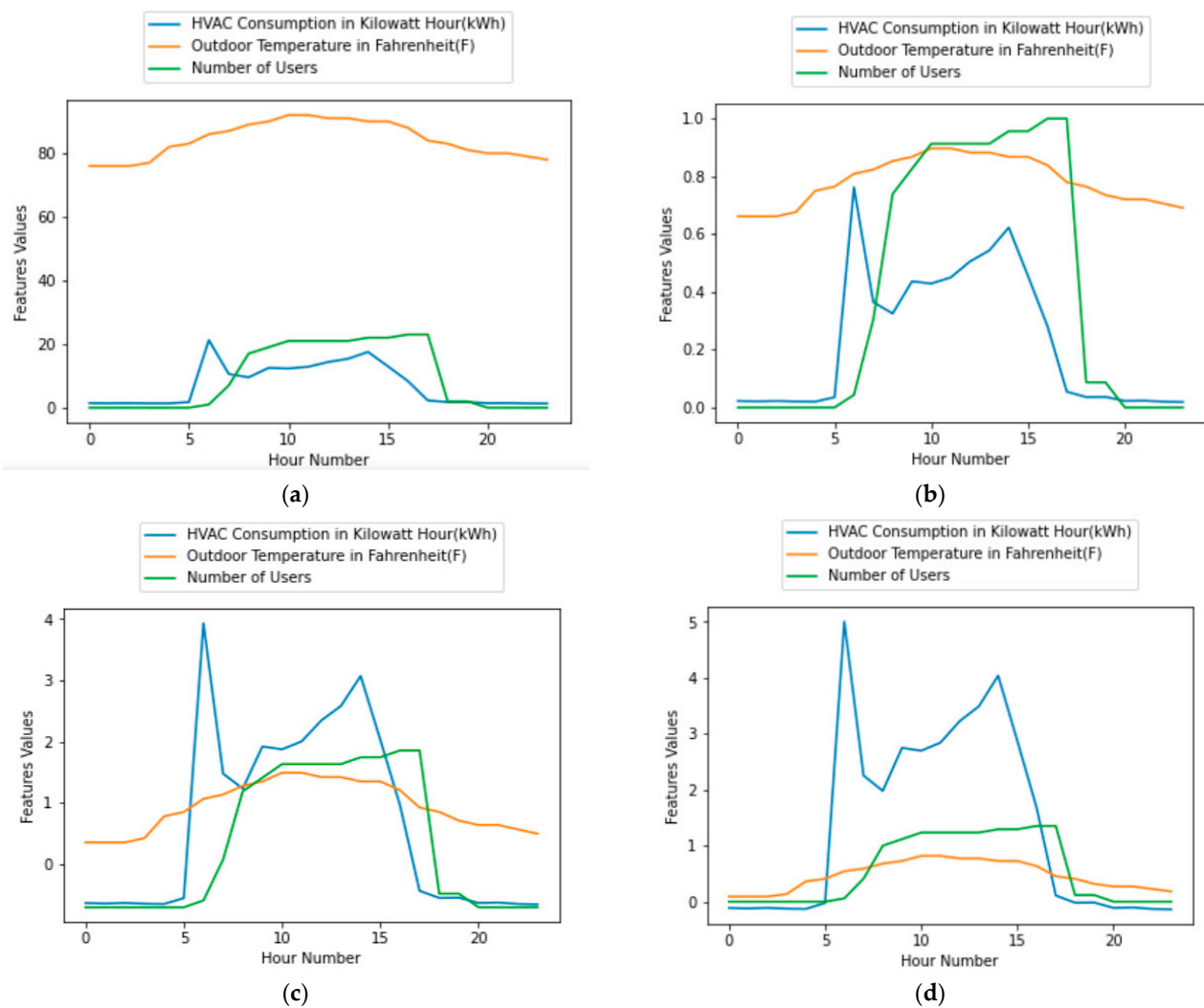


Figure 11. Effect of data scaling: (a) original data, (b) min-max scaler, (c) standard scaler, (d) robust scaler.

The next section presents the results of applying the different data processing techniques along with the K-means algorithm to identify possible HVAC system inefficiencies.

5. Applying K-Means: Results and Discussion

The K-means algorithm requires choosing an optimal number of clusters that can be identified by plotting the change of some metrics, such as elbow, silhouette, Davies Bouldin, and Calinski Harabasz (explained in Table 3), as we increase the number of clusters from 2 to N [45]. Using these metrics involves watching the change of the graph as the number of clusters increases, and then picking the number of clusters where the first elbow formation in the curve occurs. This is a common technique used in clustering, as adding another cluster beyond the elbow formation does not lead to a much better modeling of the data. Figures 12–14 present the results using the data scaled with the min-max, standard, and robust scalers.

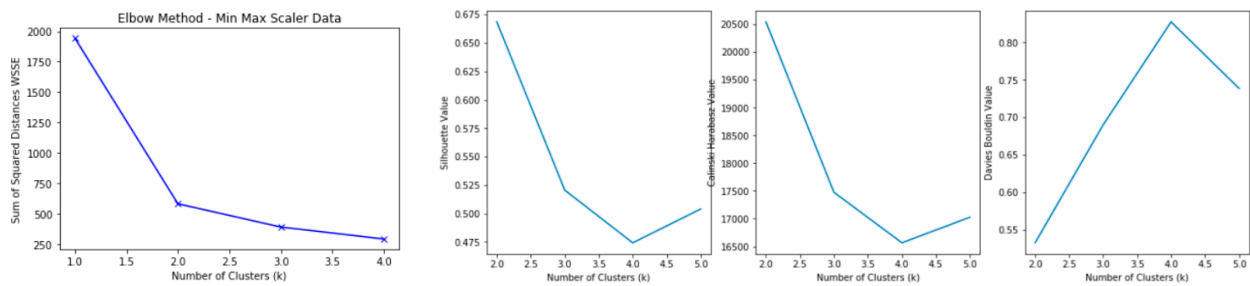


Figure 12. Metrics change for the min-max-scaled data.

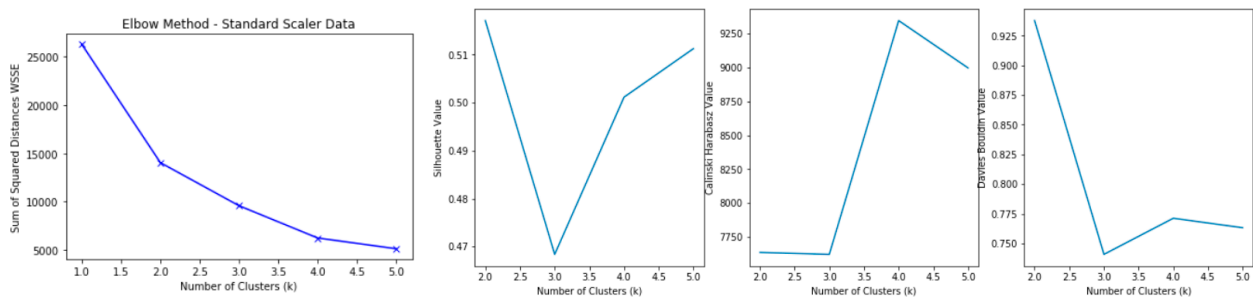


Figure 13. Metrics change for the standard-scaled data.

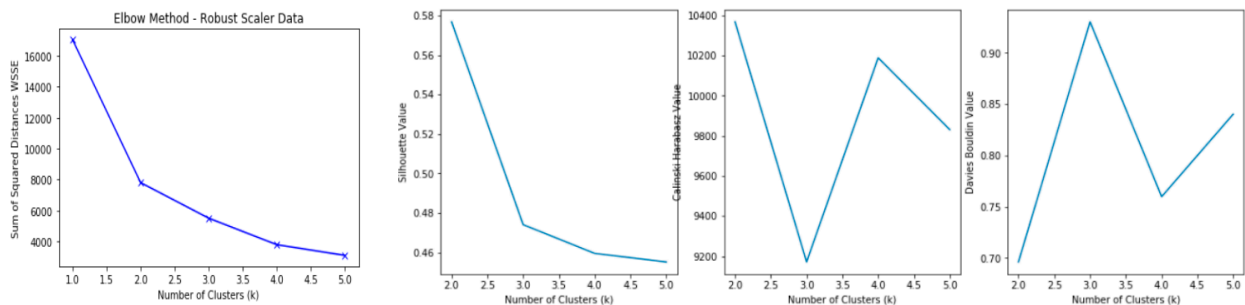


Figure 14. Metrics change for the robust-scaled data.

The elbow method showed a break at $k = 2$. However, the other metrics, including the silhouette, Davies Bouldin, and Calinski Harabasz, showed an overall elbow formation at $k = 3$ for the three scaling techniques, so we decided to use $k = 3$ as the initial number of clusters. As for the proper scaling technique, we looked at the generated clusters and evaluated the clustering method that produced the most balanced clusters. Table 5 presents the obtained results.

Table 5. Cluster distribution using different scaling methods.

Scaling Method	Clusters Size		
	Cluster 0	Cluster 1	Cluster 2
Min-Max Scaler	3775	2545	2440
Standard Scaler	3630	2569	2561
Robust Scaler	1095	5725	1940

Looking at cluster distributions, the min-max and standard scalers resulted in relatively similar distributions. However, for the robust scaler, cluster 1 contained more than 65% of the analyzed data. To get more insight on the impact of the scaling method, we decided to explore clustering using two variables: HVAC Consumption and Outdoor

Temperature, then HVAC Consumption and the Number of users. Clustering results are presented next in the next section.

5.1. Clustering HVAC Consumption and Outdoor Temperature

Table 6 presents the results of clustering the HVAC consumption along with the outdoor temperature using the three scaling methods: min-max, standard, and robust.

Table 6. Clustering the HVAC consumption and outdoor temperature: results.

Clusters Content		Min Max Scaling			Standard Scaling			Robust Scaling		
Cluster #	Number of Data Points	HVAC kWh	Outdoor Temperature (F)	Number of Data Points	HVAC kWh	Outdoor Temperature (F)	Number of Data Points	HVAC kWh	Outdoor Temperature (F)	
0	4247	2.332612	78.75936	4228	2.314971	78.85005	4143	2.254525	79.09003	
1	3167	2.961723	55.12472	3161	2.902028	55.19329	3168	2.795683	55.53346	
2	1346	12.74543	83.57281	1371	12.75034	82.94384	1449	12.5911	81.40787	

Looking at clustering membership distribution, all scaling methods resulted in a nearly similar distribution of data. Cluster 2 grouped datapoints with high-temperature values (an average >80 F), Cluster 0 assembled points where the average temperature was around 78 F, and Cluster 1 grouped data where the average temperature was around 55 F. Even though the average temperature in Cluster 1 was less than that of Cluster 0, the average HVAC consumption in cluster #1 was higher than the value in Cluster 0. To further analyze the data, we evaluated the timeslots from cluster #1, highlighted in green, where the HVAC consumption was high while the outdoor temperature was low. The analysis produced an average of 1340 timeslots (for every scaling method) where the consumption was more than 2.33 kWh when the outdoor temperature was low. Table 7 presents the timeslots corresponding to every month (for data scaled used the min-max scaler).

Table 7. Number of timeslots/month with high HVAC consumption.

Month	Oct_18	Nov_18	Dec_18	Jan_19	Feb_19	Mar_19	Apr_19	May_19
Number of timeslots	254	325	85	319	177	114	70	2

Given the length of the obtained results, we present a screenshot of a sample of some concerned months (columns represent the date, hour, HVAC consumption, and outdoor temperature) in Figure 15.

2018-10-18	6	5.86489	58	2019-02-04	6	6.568737	66
	7	4.142852	59		7	5.375346	66
	8	6.683803	59		8	4.760542	67
	9	7.540262	60		9	4.86422	67
	10	7.374827	62				
	11	7.896883	63				
2018-12-22	2	2.684848	47	2018-11-20	0	2.529073	50
	5	2.731883	48		1	3.033046	51
2019-01-22	6	6.085206	63		3	3.606199	49
	7	2.565032	63		4	4.424087	48
	8	4.602202	63		6	2.818166	48
	9	4.173711	65		7	6.372232	47
	10	4.341053	67		8	5.276981	49
			9		4.955863	51	
			10		5.10608	54	
			11		4.834653	57	
			12		5.177856	59	

Figure 15. A sample of days with high HVAC consumption when the outdoor temperature was low.

Using the clustering results, it was possible to identify timeslots where the HVAC consumption was high while the outdoor temperature was low. Table 8 presents the hours and frequency where such status occurred.

Table 8. Timeslots with high HVAC consumption despite low temperatures.

Hour	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Freq	18	22	17	20	29	32	85	112	112	113	105	94	88	79	71	73	73	76	37	25	18	13	14	14

Excluding the timeslots where the setpoint is changed, i.e., 6:00 a.m. and 6:00 p.m., other slots between 7:00 and 16:00, highlighted in green, are important for building managers to consider as they present potential wastes of energy in the HVAC system.

5.2. Clustering HVAC Consumption and Number of Users

In this section, we explore the results of clustering the HVAC consumption along with the number of users. As the number of users, between 6:00 a.m. and 6:00 p.m., was based on the assumption that every person remained in the building until 5:30 p.m., this assumption may not have been 100% correct. Thus, we limited our focus to timeslots where the number of occupants was 0, but the HVAC consumption had a high value. Table 9 presents the clustering results.

Table 9. Clustering the HVAC consumption and the number of users: results.

Clusters Content	Min Max Scaling			Standard Scaling			Robust Scaling		
	Cluster #	Number of Data Points	HVAC kWh	N_Users	Number of Data Points	HVAC kWh	N_Users	Number of Data Points	HVAC kWh
0	5973	1.988811	0.693286	1711	5.62081	20.877265	5716	1.791676	0.621064
1	2440	7.964945	20.47541	6029	2.051677	0.713717	1984	5.705946	17.899194
2	347	14.778369	2.951009	1020	14.17153	14.804902	1060	14.037649	15.153774

By clustering the HVAC consumption and number of users, we noted that Cluster 2, highlighted in green, grouped datapoints where the average consumption was about 14.8 kWh when the number of users was relatively small. Some important insights and inefficiencies were concluded after analyzing Cluster 2 datapoints. The results of the analysis are summarized in Table 10.

Table 10. Identifying the timeslots with inefficiencies.

Timeslots	Increase in HVAC Consumption: Explanation
All working days 6:00 to 7:00 a.m. Some 6:00 a.m. to 7:00 a.m. slots for vacations still have a high consumption Example: 23,24 November; 31 December 2018, 21 January 2019, 18 February	Setpoint was changed
- Columbus Day (Monday: 8 October 2018): 06:00 a.m. to 17:00 - Thanksgiving Day (Thursday: 22 November 2018) - Saturday 10 August 2019: high consumption from 7:00 a.m. to 17:00	Some areas did not switch to unoccupied mode properly in the BAS system, causing an inefficiency Either HARC staff worked during this day, or the day was not input as a vacation day in the system, causing an inefficiency
November 24, 2018: 3:00 a.m.: 11 kWh 26 November: 2:00 a.m. and 5:00 a.m.: more than 9 kWh	Either an outlier or inefficiency. The building had a minimum temperature setpoint of 65F. When a zone reached that temperature during winter nights, the heating systems warmed up that specific zone.
For all working days between 6:00 a.m. to 17:00	The control system did not allow the configuration of occupancy-based setpoints

Looking at the different timeslots grouped in Cluster 2, we identified the timeslots that had a high consumption when the number of users was zero. Table 11 shows the results.

Table 11. Timeslots with high HVAC consumption when the number of users is low.

Hour	0	2	3	4	5	6	7	13	14	15	16	17	18	20
Freq	1	1	1	1	1	8	2	4	3	3	4	4	2	1

To estimate the energy waste, we explored the timeslots grouped in Cluster 0 and Cluster 2, respectively. The former aggregated timeslots where the HVAC consumption and number of users were both of low values, while the latter is the cluster where inefficiencies were found. We computed the average consumption of every timeslot in Cluster 0, then compared the values to the corresponding values grouped in Cluster 2 (same timeslot, same temperature). Our analysis showed that a minimum of 1.87% savings could be achieved for 1-year consumption if the HVAC system was configured correctly by accounting for the number of users in the building.

Finally, it was important to look at clustering results when considering the three features (the HVAC consumption, number of users, and outdoor temperature). The results are presented in Table 12. The goal is to identify high average power consumption when the outdoor temperature and the number of people values are low. This provides an indication to the building managers to change the setting of these conditions to a more energy-efficient condition, such as lowering the setpoint of the thermostat.

Table 12. Clustering the HVAC consumption, number of users, and outdoor temperature: results.

Clusters Content	Min Max Scaler			Standard Scaler			Robust Scaler		
	HVAC kWh	Outdoor Temperature (F)	N_Users	HVAC kWh	Outdoor Temperature (F)	N_Users	HVAC kWh	Outdoor Temperature (F)	N_Users
Cluster 0	2.92041	79.321325	0.79497	1.98926	79.047934	0.625069	13.87707	83.121461	15.52785
Cluster 1	7.96495	74.409836	20.4754	8.98954	76.07863	18.82016	1.774806	70.057293	0.747598
Cluster 2	2.35077	55.230648	0.8503	2.39227	54.342054	1.759859	5.714216	66.734021	17.44433

All three scaling methods produced good clustering results regarding the outdoor temperature and number of users. For every scaling method used, the K-means algorithm resulted in a cluster with high consumption proportional to the high temperature and number of users. In addition, the results showed a second cluster grouping data with the high temperature and low number of users, as well as a third cluster grouping data with a low consumption proportional to the low temperature and low number of users. Compared to the previous results, the second cluster revealed good results, as we could identify the timeslots with energy losses which could be considered for future HVAC settings.

Besides accounting for occupancy, we explored the energy savings that could be achieved if we delayed the change of the setpoint from 06:00 a.m. to 07:00 a.m., given that the number of occupants is only one. Our analysis showed that an extra savings of 4% could be achieved if the morning setpoint change was shifted by an hour, which is another suggestion that the energy manager can consider for future HVAC settings.

6. Conclusions and Future Work

HVAC systems consume a significant amount of energy in a building, which is particularly apparent in hot and humid areas such as Houston, TX, USA. High temperatures lead to an increased demand for cooling using HVAC systems throughout the year. Given the improvement of information and communication technology (ICTs), significant attention is devoted to collecting building-related data that analysts can use to uncover energy inefficiencies. This paper presents a step-by-step methodology able to unveil improvement opportunities in HVAC management using the K-means algorithm, an unsupervised machine learning algorithm, in highly efficient buildings. The data used in this paper were collected from a weather station, building access cards, and a meter. The meter represents the 1-year consumption of a LEED Platinum-, Energy Star 99/100-, and Net-Zero-certified office building owned and operated by the Houston Advanced Research Center (HARC).

The results show that using time series analysis and an unsupervised learning technique can distinguish timeslots to reduce or eliminate the use of HVAC consumption. Our study proves the possibility of saving up to 6% in energy that energy managers can consider for the future configuration of the HVAC system.

For future work, we propose an implementation of a real-time collection of the number of users in the building and measuring the inside temperature, as this is important to understand the relationship between the indoor temperature and the changes in HVAC consumption. Storing the setpoint used is another important future action, as improving the setpoint affects the HVAC system consumption in the next hour(s) [46]. In addition to collecting more building-related data, our next research work will focus on comparing the performance of the K-means algorithm to other clustering techniques in identifying energy inefficiencies, as this is an important step to identify the best clustering technique to use for the problem addressed in this paper.

Author Contributions: H.T. worked on Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Visualization, and Writing—Original Draft. D.B. worked on Supervision, Project Administration, Conceptualization, Methodology, Investigation, and Writing—Review & Editing. H.B. worked on Conceptualization and Investigation. C.G. worked on providing Resources, Conceptualization, Investigation, Validation, and Writing—Review & Editing. M.E. worked on Supervision and Writing—Review & Editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted according to the guidelines to the declaration of Helsinki and approved by the Institutional Review Board of the University of Houston, IRB ID: MOD00000338 approved on 26 April 2017, and valid until 5 October 2021.

Informed Consent Statement: Not Applicable.

Data Availability Statement: The data analyzed in this paper cannot be shared publicly, as an NDA was signed by Houston Advance Research Center, Houston University, and Al Akhawayn University to use the data for this research only.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Air Conditioning. Department of Energy. Available online: <https://www.energy.gov/energysaver/home-cooling-systems/air-conditioning> (accessed on 11 September 2020).
2. Javed, A.; Larijani, H.; Wixted, A. Improving Energy Consumption of a Commercial Building with IoT and Machine Learning. *IT Prof.* **2018**, *20*, 30–38. [CrossRef]
3. ENERGY STAR Certification for Your Building. ENERGY STAR Buildings and Plants. ENERGY STAR. Available online: <https://www.energystar.gov/buildings/facility-owners-and-managers/existing-buildings/earn-recognition/energy-star-certification> (accessed on 18 January 2021).
4. Talei, H. Smart Campus Energy Management System: Advantages, Architectures, and the Impact of Using Cloud Computing. In Proceedings of the ICSDE '17, Rabat, Morocco, 21–23 July 2017.
5. Ardiyanto, D.; Pipattanasomporn, M.; Rahman, S.; Hariyanto, N. Suwarno Occupant-based HVAC Set Point Interventions for Energy Savings in Buildings. In Proceedings of the 2018 International Conference and Utility Exhibition on Green Energy for Sustainable Development (ICUE), Phuket, Thailand, 24–26 October 2018; pp. 1–6.
6. Capozzoli, A.; Piscitelli, M.S.; Gorrino, A.; Ballarini, I.; Corrado, V. Data analytics for occupancy pattern learning to reduce the energy consumption of HVAC systems in office buildings. *Sustain. Cities Soc.* **2017**, *35*, 191–208. [CrossRef]
7. Lu, J.; Sookoor, T.; Srinivasan, V.; Gao, G.; Holben, B.; Stankovic, J.; Field, E.; Whitehouse, K. The smart thermostat. In Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems—SenSys '10, Delft, The Netherlands, 6–8 November 2010; pp. 211–224.
8. Fan, C.; Xiao, F.; Wang, S. Development of prediction models for next-day building energy consumption and peak power demand using data mining techniques. *Appl. Energy* **2014**, *127*, 1–10. [CrossRef]
9. Fan, C.; Xiao, F.; Li, Z.; Wang, J. Unsupervised data analytics in mining big building operational data for energy efficiency enhancement: A review. *Energy Build.* **2018**, *159*, 296–308. [CrossRef]

10. Gaitani, N.; Lehmann, C.; Santamouris, M.; Mihalakakou, G.; Patargias, P. Using principal component and cluster analysis in the heating evaluation of the school building sector. *Appl. Energy* **2010**, *87*, 2079–2086. [[CrossRef](#)]
11. Wall, J.; Guo, Y.; Li, J.; West, S. A Dynamic Machine Learning-Based Technique for Automated Fault Detection in HVAC Systems. *ASHRAE Trans.* **2011**, *117*, 449–456.
12. Yu, F.; Chan, K. Assessment of operating performance of chiller systems using cluster analysis. *Int. J. Therm. Sci.* **2012**, *53*, 148–155. [[CrossRef](#)]
13. Fan, C.; Xiao, F.; Yan, C. A framework for knowledge discovery in massive building automation data and its application in building diagnostics. *Autom. Constr.* **2015**, *50*, 81–90. [[CrossRef](#)]
14. Howard, P.; Runger, G.; Reddy, T.A.; Katipamula, S. Automated Data Mining Methods for Identifying Energy Efficiency Opportunities Using Whole-Building Electricity Data. *ASHRAE Trans.* **2016**, *122*, 422–433.
15. Yu, F.; Chan, K. Using cluster and multivariate analyses to appraise the operating performance of a chiller system serving an institutional building. *Energy Build.* **2012**, *44*, 104–113. [[CrossRef](#)]
16. Li, G.; Hu, Y.; Chen, H.; Li, H.; Hu, M.; Guo, Y.; Liu, J.; Sun, S.; Sun, M. Data partitioning and association mining for identifying VRF energy consumption patterns under various part loads and refrigerant charge conditions. *Appl. Energy* **2017**, *185*, 846–861. [[CrossRef](#)]
17. Cerquitelli, T.; Chicco, G.; Di Corso, E.; Ventura, F.; Montesano, G.; Del Pizzo, A.; Gonzalez, A.M.; Sobrino, E.M. Discovering electricity consumption over time for residential consumers through cluster analysis. In Proceedings of the 2018 International Conference on Development and Application Systems (DAS), Suceava, Romania, 24–26 May 2018; pp. 164–169. [[CrossRef](#)]
18. Tang, F.; Kusiak, A.; Wei, X. Modeling and short-term prediction of HVAC system with a clustering algorithm. *Energy Build.* **2014**, *82*, 310–321. [[CrossRef](#)]
19. Xiao, F.; Fan, C. Data mining in building automation system for improving building operational performance. *Energy Build.* **2014**, *75*, 109–118. [[CrossRef](#)]
20. Ren, X.; Yan, D.; Hong, T. Data mining of space heating system performance in affordable housing. *Build. Environ.* **2015**, *89*, 1–13. [[CrossRef](#)]
21. Dobiáš, J.; Macek, D. Leadership in Energy and Environmental Design (LEED) and its Impact on Building Operational Expenditures. *Procedia Eng.* **2014**, *85*, 132–139. [[CrossRef](#)]
22. ASHRAE. *Thermal Environmental Conditions for Human Occupancy (ANSI/ASHRAE 55-2017)*; American Society of Heating, Refrigeration and Air-Conditioning Engineers: Atlanta, GA, USA, 2017; Available online: <https://www.ashrae.org/technical-resources/standards-and-guidelines> (accessed on 15 August 2021).
23. Anderson, A.; Rezaie, B. Geothermal technology: Trends and potential role in a sustainable future. *Appl. Energy* **2019**, *248*, 18–34. [[CrossRef](#)]
24. Air Conditioner Basics Part II—Thermodynamics. Available online: <https://sandium.com/general-hvac/air-conditioner-basics-part-ii-thermodynamics.html> (accessed on 1 August 2021).
25. Serale, G.; Fiorentini, M.; Capozzoli, A.; Bernardini, D.; Bemporad, A. Model Predictive Control (MPC) for Enhancing Building and HVAC System Energy Efficiency: Problem Formulation, Applications and Opportunities. *Energies* **2018**, *11*, 631. [[CrossRef](#)]
26. Li, W.; Koo, C.; Hong, T.; Oh, J.; Cha, S.H.; Wang, S. A novel operation approach for the energy efficiency improvement of the HVAC system in office spaces through real-time big data analytics. *Renew. Sustain. Energy Rev.* **2020**, *127*, 109885. [[CrossRef](#)]
27. Papadopoulos, S.; Kontokosta, C.; Vlachokostas, A.; Azar, E. Rethinking HVAC temperature setpoints in commercial buildings: The potential for zero-cost energy savings and comfort improvement in different climates. *Build. Environ.* **2019**, *155*, 350–359. [[CrossRef](#)]
28. Do, H.; Cetin, K.S. Data-Driven Evaluation of Residential HVAC System Efficiency Using Energy and Environmental Data. *Energies* **2019**, *12*, 188. [[CrossRef](#)]
29. Jung, W.; Jazizadeh, F. Human-in-the-loop HVAC operations: A quantitative review on occupancy, comfort, and energy-efficiency dimensions. *Appl. Energy* **2019**, *239*, 1471–1508. [[CrossRef](#)]
30. Erickson, V.L.; Carreira-Perpiñán, M.Á.; Cerpa, A.E. Occupancy Modeling and Prediction for Building Energy Management. *ACM Trans. Sens. Networks* **2014**, *10*, 1–28. [[CrossRef](#)]
31. Lawton, R. Time Series Analysis and its Applications. *Int. J. Forecast.* **2001**, *17*, 299–301. [[CrossRef](#)]
32. Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. Optimal Filters in the Time Domain. *Natur. Comput. Ser.* **2009**, *2*, 1–18. [[CrossRef](#)]
33. Talei, H.; Essaïdi, M.; Benhaddou, D. An End to End Real Time Architecture for Analyzing and Clustering Time Series Data: Case of an Energy Management System. In Proceedings of the 2018 6th International Renewable and Sustainable Energy Conference (IRSEC), Rabat, Morocco, 5–8 December 2018; pp. 1–7.
34. Tureczek, A.; Nielsen, P.S.; Madsen, H. Electricity Consumption Clustering Using Smart Meter Data. *Energies* **2018**, *11*, 859. [[CrossRef](#)]
35. Tureczek, A.M.; Nielsen, P.S.; Madsen, H.; Brun, A. Clustering district heat exchange stations using smart meter consumption data. *Energy Build.* **2019**, *182*, 144–158. [[CrossRef](#)]
36. Bahmani, B.; Kumar, R.; Vassilvitskii, S. Scalable K-Means++. *ArXiv* **2012**, arXiv:1203.64025, 622–633. [[CrossRef](#)]
37. Elbow Method—Yellowbrick v1.1. Documentation. Available online: <https://www.scikit-yb.org/en/latest/api/cluster/elbow.html> (accessed on 1 October 2020).

38. Sklearn.metrics.silhouette_score—Scikit-Learn 0.23.2. Documentation. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html (accessed on 1 October 2020).
39. Sklearn.metrics.davies_bouldin_score—Scikit-Learn 0.23.2. Documentation. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html (accessed on 1 October 2020).
40. Sklearn.metrics.calinski_harabasz_score—Scikit-Learn 0.23.2. Documentation. Available online: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.calinski_harabasz_score.html (accessed on 1 October 2020).
41. Nepal, B.; Yamaha, M.; Sahashi, H.; Yokoe, A. Analysis of Building Electricity Use Pattern Using K-Means Clustering Algorithm by Determination of Better Initial Centroids and Number of Clusters. *Energies* **2019**, *12*, 2451. [[CrossRef](#)]
42. Sklearn.preprocessing.MinMaxScaler—Scikit-Learn 0.23.2. Documentation. Available online: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html> (accessed on 1 October 2020).
43. Sklearn.preprocessing.StandardScaler—Scikit-Learn 0.23.2. Documentation. Available online: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html> (accessed on 1 October 2020).
44. Sklearn.preprocessing.RobustScaler—Scikit-Learn 0.23.2. Documentation. Available online: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html> (accessed on 1 October 2020).
45. Yuan, C.; Yang, H. Research on K-Value Selection Method of K-Means Clustering Algorithm. *J* **2019**, *2*, 226–235. [[CrossRef](#)]
46. Wang, N.; Zhang, J.; Xia, X. Energy consumption of air conditioners at different temperature set points. In Proceedings of the IEEE Africon 11, Victoria Falls, Zambia, 13–15 September 2011; pp. 1–6. [[CrossRef](#)]