Energy Consumption Pattern Discovery Using Clustering and Association Rule Mining

Alex Jurcich

Abstract

As smart energy metering expands, apartment energy data has become important for understanding and improving energy efficiency. My project aims to apply modern data techniques to identify patterns in daily electricity use across apartments located at UMASS. Using this dataset, which provides 3 years of electricity and weather records, the objective is to uncover behavioral clusters and learn relationships between environmental conditions and energy demands. I will combine different clustering algorithms with association rule mining to generate interpretable insights into apartments' energy usage. Performing this analysis can help develop energy usage patterns and understand how the seasons relate to energy usage.

Keywords:

Electricity Consumption, Smart Apartments, Clustering, Gaussian Mixture Models, DBSCAN, PCA, Association Rule Mining, Apriori, FP-Growth

Introduction

The increase in smart meter data has transformed the ability of data professionals to analyze the behavior of energy usage. Numerous factors, such as temperature, season, occupancy, and appliance usage, influence apartment-level energy consumption. These relationships are often nonlinear and difficult to pinpoint using summary statistics. Data Mining allows the use of powerful tools for discovering hidden patterns and structure within a dataset.

This project focuses on applying clustering and association rule mining to multiple years of UMASS apartment electricity data records. Unlike many other energy studies that examine household totals or forecast future energy consumption, this project will try to identify behavioral patterns, such as group apartments with similar daily energy consumption/patterns, and combinations of environmental factors that shift energy consumption. I will use clustering algorithms such as K-means, DBSCAN, and Gaussian Mixture Models to segment apartments. I will also use association rule mining with the intent to reveal co-occurring conditions such as temperature, humidity, and total energy consumption. Altogether, these techniques can provide actionable insights and explain why electricity usage varies. This can lead to supporting informed energy management and planning.

2 Methods

The dataset used is the **Smart* Data Set for Sustainability** from UMASS Trace Repository, containing daily electricity usage for 114 single-family apartments for the period 2014-2016, as well as weather data. This project will proceed in three main stages:

1. **Data Preprocessing:** For the preprocessing of this dataset, I will perform it in three ordered steps:
    a. Data Cleaning, where I will combine over 300 .csv files, parse and convert timestamps to daily summaries (mean, min, max, standard deviation, total sum), remove missing data and outliers, and apply IQR-based filtering to reduce the extremes in the dataset.
    b. Feature Engineering, where I will create categorical variables for the season (winter, spring, summer, fall) and weekend vs weekday.
    c. Dimensionality Reduction, where I will standardize numerical features and apply PCA to be able to visualize the clustering results.

2. **Cluster Analysis:** Apply unsupervised methods, including K-means, DBSCAN, and GMM, to identify natural groupings amongst apartments. Clustering will be evaluated using the appropriate metrics depending on the method, as well as visually.

3. **Association Rule Mining:** Perform the Apriori and FP-Growth algorithms to uncover co-occurring patterns between energy usage and environmental conditions.

Results

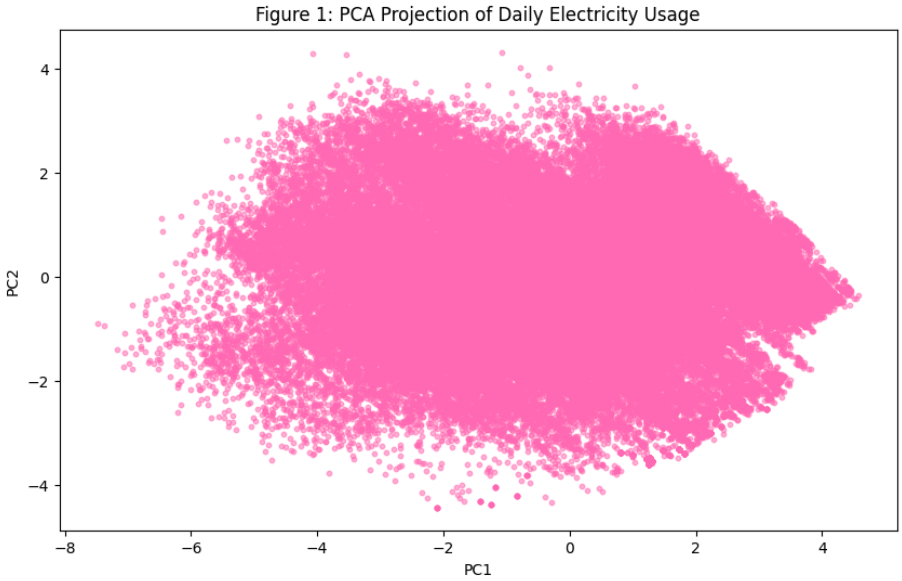
1 Data Overview and Cleaning Outcomes

The dataset comprised 90,348 observations across 114 apartments and spanned three years (2014-2016). After merging the electricity and weather files, cleaning and removing rows, the dataset was reduced to 88,303 observations. After inspecting the distributions of the variables, some outliers would have made clustering impossible, so I performed IQR-based removal on total power, mean power, maximum power, minimum power, and standard deviation. After this removal, the final result was 69,603 daily apartment records.

2 Principal Component Analysis

Principal Component Analysis was applied to assess the underlying structure and evaluate how well different features separate the daily observations. Before performing PCA, I made sure to standardize the data using **StandardScaler** on all the numeric variables. After

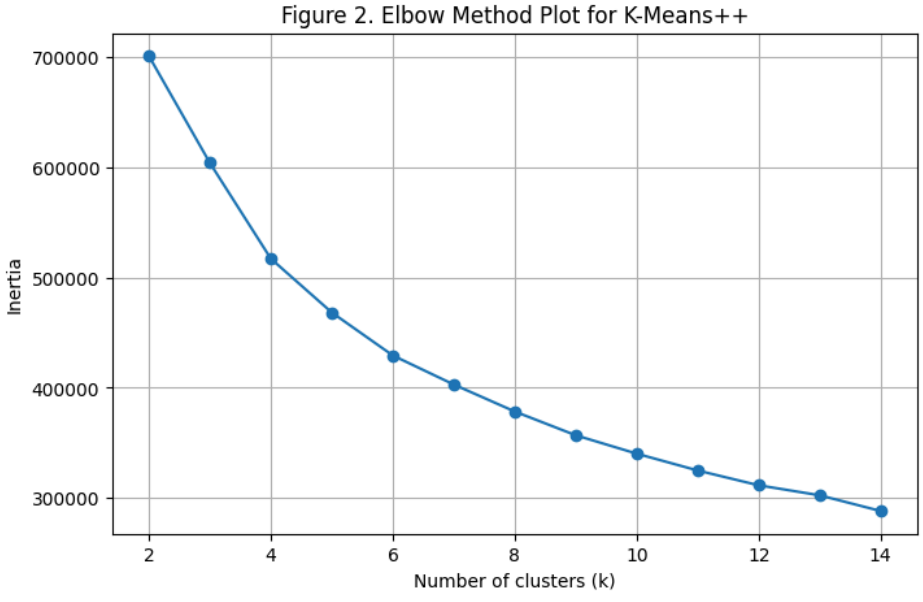
ensuring PCA maintains 90% variance, it showed the first 3 components captured 63% of the total variance (PC1 = 38.1%, PC2 = 13.8%, PC3 = 11.2%). PC1 was associated with temperature and total power, and PC2 showed patterns in humidity and weekend/weekday effects. **Figure 1** shows the dataset projected into the first two principal components.



The PCA projection shows the dataset is extremely dense. All observations are tight in one cluster with little to no separation. The figure will help explain the later results, where clustering struggled due to the sheer density of the data.

3 K-Means++ Clustering Results

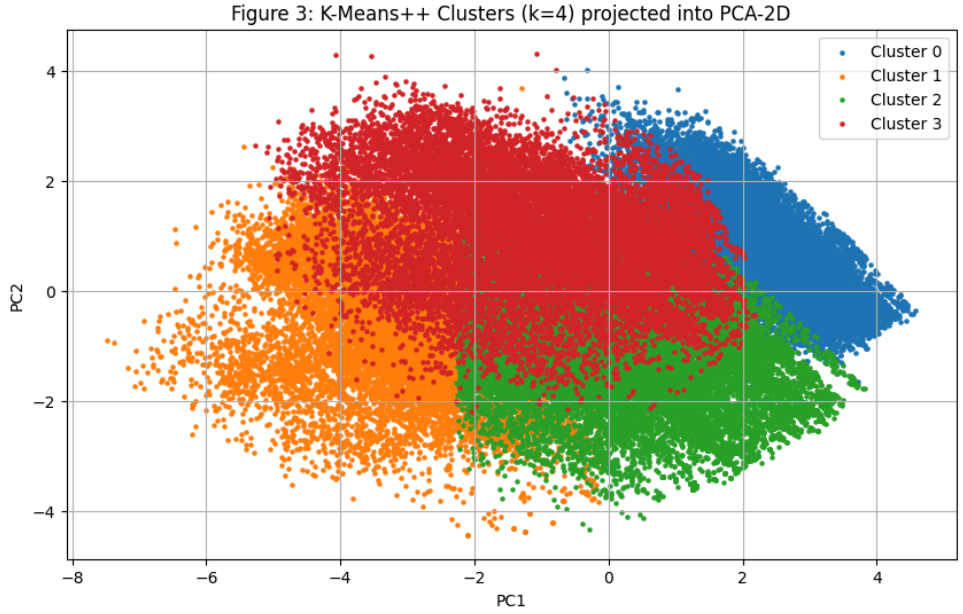
I selected K-Means++as the first clustering method because it's one of the most basic and well-known clustering methods. Prior to performing K-Means, I had to find the proper number of clusters. To do this, I performed the elbow methods with inertia on the y-axis. As you can see in Figure 2, inertia began to flatten at k = 4.



Interia decreases steeply until k = 4. This indicates that 4 clusters are the most realistic and effective way to cluster the data.

Next, I applied K-Means to the standardized dataset. Then I visualized the clusters in our predefined PCA space to assess their structure. As you can see in Figure 3, the clusters show significant overlap, which was expected due to the density of the original PCA plot (Figure 1). Despite the overlap, we can see that 4 areas are beginning to form.

Figure 3: K-Means++ Clusters (k=4) projected into PCA-2D



The four K-Means clusters overlap in PCA space, reflecting the sheer density of the dataset.

Since it's difficult to interpret the clusters visually, a better option is to review the average daily metrics for mean power, temperature, total power, and humidity for each cluster. Figure 4 summarizes those metrics.

Figure 4: K-Means Cluster Summary Statistics

cluster_k4	avg_total_power	avg_mean_power	avg_temp	avg_humidity	avg_max_temp	avg_min_temp	count
0	326.54	0.45	70.06	0.70	80.55	60.25	21705
1	1256.18	1.66	33.53	0.70	41.39	26.92	13066
2	417.85	0.71	54.30	0.73	64.59	45.80	19240
3	890.38	1.11	49.32	0.59	60.36	39.57	15592

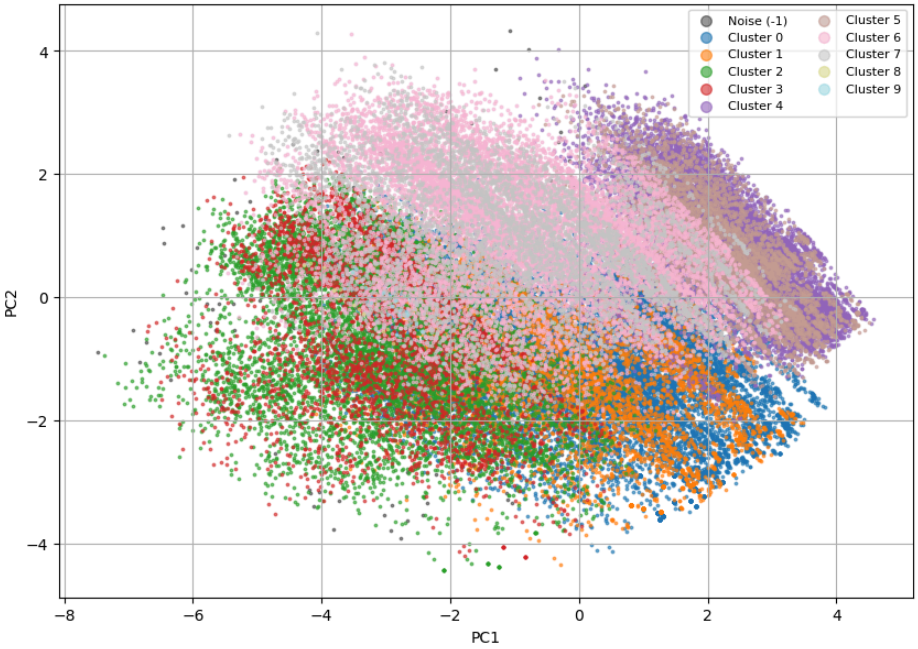
Summary statistics reveal that Cluster 1 contains cold days with the highest energy usage, Cluster 2 contains hot days, and Clusters 2 & 3 represent transitional seasons.

Taking in all of the results from K-Means, it seems that temperature and season are the main factors influencing apartment-level energy consumption. Although the clusters overlap in PCA space, the numerical summaries show clear separation in energy usage. Given the density of the data, the next step is to use a density-based clustering method and see if it can undercover any insight.

4 DBSCAN

To ensure that DBSCAN would find any structure within the data, I conducted a grid search across two DBSCAN parameters, **eps** (0.2-1.0) and **min_samples** (10, 20, 30). This allowed DBSCAN to be evaluated with both tight parameters and more permissive thresholds. The best result was an **eps** = 1.0 and **min_samples** = 10. These parameter values produced the lowest noise rates, making it the best possible choice for inspecting DBSCAN.

Figure 5: DBSCAN Cluster Assignments in PCA Space



Even after using the best possible DBSCAN settings, the algorithm generates widespread noise and many small clusters. This is because our dataset is so dense that it struggles to find anything meaningful.

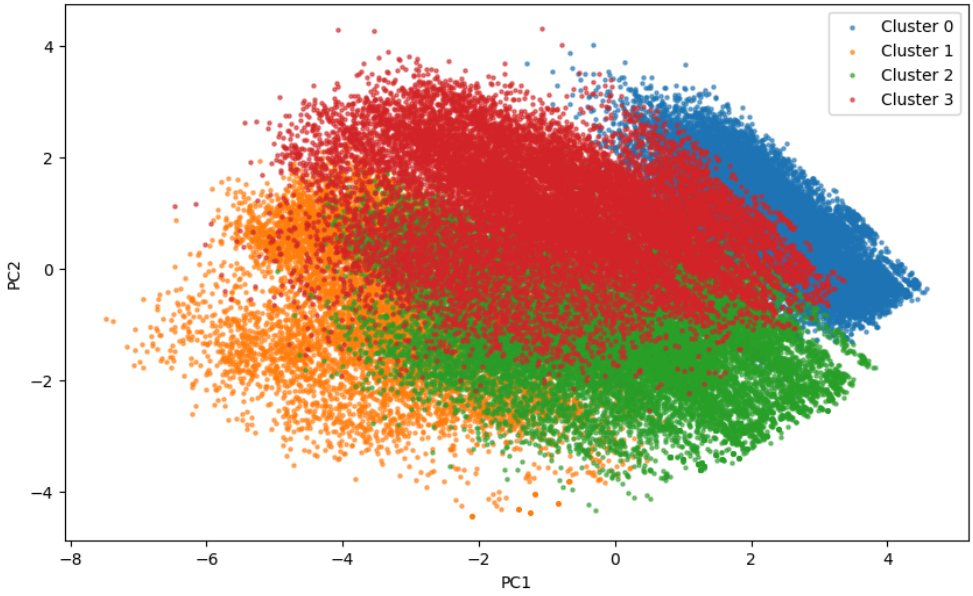
This outcome from DBSCAN and some research shows the structure of the electricity data itself. Energy data tends to form dense clouds without any low-density valleys. Since DBSCAN relies on gaps, it is not well-suited for this dataset. As a result, DSCAN fails to provide anything meaningful segmentation.

5 Gaussian Mixture Models (GMM)

Given that K-Means and DBSCAN didn't visually showcase anything meaningful, my next step was to apply GMM. Unlike K-Means, which forces hard clusters. GMM models each cluster as a Gaussian distribution and allows for soft assignments. GMM most likely will not be able to show any separation, but since it's a different method, it is worth exploring.

To select the number of components, I fit GMM with k = 2 through 8 and compared the AIC and BIC. Like before, I will be looking for the elbow of the curves to pick the best k. The result was very similar to the previous elbow plot we developed for K-Means (Figure 2). Thus, k = 4 was the best possible choice, due to there being 4 seasons.

Figure 6: GMM Clusters in PCA Space



GMM produced an almost identical result to K-Means with four overlapping clusters when projected into PCA space.

Even though GMM and K-Means use very different methods to separate the data, they both produced nearly identical results. This is because temperature and season are the main drivers when it comes to energy consumption. Overall, GMM didn't visually reveal any new structure within the data. However, to compare GMM to K-Means, we must compare the average metrics of each cluster.

Figure 7: GMM Cluster Summary Statistics

cluster_gmm	avg_total_power	avg_mean_power	avg_temp	avg_humidity	count
0	341.03	0.47	70.05	0.70	20600
1	1151.70	1.59	33.21	0.70	10670
2	561.11	0.84	52.41	0.73	21437
3	833.36	1.07	50.42	0.60	16896

Average temperature, humidity, and energy metrics are nearly identical between GMM and K-Means clusters. Both methods are dominated by seasonal energy use.

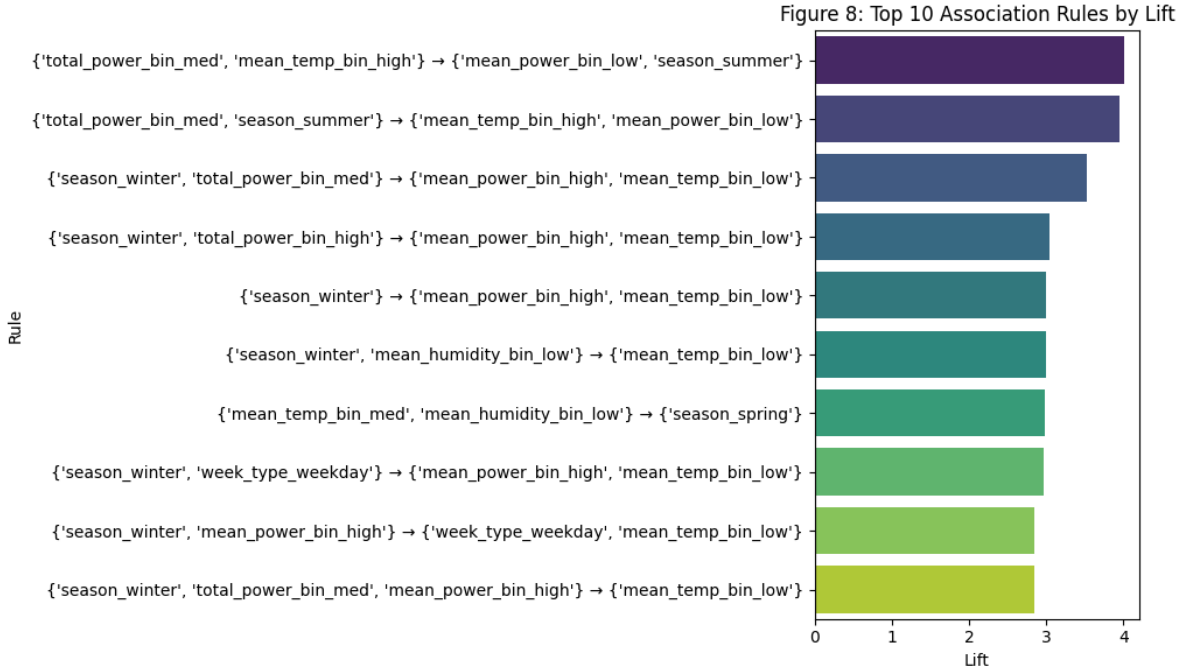
Without proper investigation, GMM should theoretically outperform K-means when clusters overlap. However, we now know that the dataset is dominated by temperature. Which makes sense when thinking of energy usage, especially in a climate with 4 seasons. Thus, after clustering our data, we now know that electricity usage varies with environmental conditions. Both K-Means and GMM confirm this.

6 Association Rule Mining

To add to the clustering analysis, association rule mining was used to identify any co-occurring environmental and energy-usage conditions. Performing this answers questions like: Whether certain temperature ranges are associated with high or low electricity usage. To perform this, I used two methods: Apriori and FP-Growth. To make the results interpretable, I converted the numeric variables into categorical bins using quantile-based binning.

After performing both methods, they each produced identical results. The rules that were found were not surprising after performing the cluster analysis. Temperature and season were the dominant drivers in energy usage. For example, days with **high total power** almost always co-occurred with **low mean temperature**. Another example was that days with **low total power** were commonly associated with **mild temperatures**.

Figure 8: Top 10 Association Rules by Lift



Frequent itemsets reflect environmental conditions. Days with high power usage coincided with extreme temperatures, while low usage days occurred in mild temperatures.

Unsurprisingly, rules that involved weekday/weekend, humidity, or seasonality had much lower lift values. Thus proving that temperature is the main influence on energy consumption. Once again showing that the strongest patterns are environmental, not behavioral. While association rule mining didn't reveal any new information, it was a good validation step for what was inferred from the clustering results.

Discussion

1 Conclusion

The goal of this project was to identify patterns in apartment electricity usage and determine the factors that contribute to variation in daily demand. Using multiple clustering methods and association rule mining showed that temperature is the dominant driver of energy usage. When apartments use their heaters, they require significantly more energy, resulting in higher overall consumption. I was initially surprised that the data did not show greater usage in the summer months, but after researching, most homes and apartments use far more energy for heating than for cooling. Additionally, some apartments may not have cooling systems at all, which could also explain the lower averages observed during the summer.

2 Limitations

The analysis I performed is not perfect and has limitations that should be acknowledged. First, using daily aggregated data can smooth out intra-day patterns, limiting the ability to reveal key behavioral differences between day and night. Second, the data comes from single-family homes and doesn't include any information on occupancy or appliances. Which makes it difficult to reveal the resident difference. The third limitation is that all the apartments were located at UMASS, thus the findings of this analysis don't generalize to other regions. The last limitation is that temperature masks almost all the factors, making it difficult to see if they have an impact or not.

References

[1] Smart*: An Open Data Set and Tools for Enabling Research in Sustainable Homes. Sean Barker, Aditya Mishra, David Irwin, Emmanuel Cecchet, Prashant Shenoy, and Jeannie Albrecht. Proceedings of the 2012 Workshop on Data Mining Applications in Sustainability (SustKDD 2012), Beijing, China, August 2012. PDF Smart* | UMass Trace Repository

[2] Tan, P., Steinbach, M., Karpatne, A., & Kumar, V. *Introduction to Data Mining (2nd ed.)*. Available via the U. Minnesota website. College of Science and Engineering

[3] "MMDS" (Massive Data Mining Series) – http://www.mmds.org/

[4] Course slides from CSCI-B 565 (Data Mining)

[5] U.S. Energy Information Administration. (2023, June 15). *Space heating consumed the most energy of any end use in homes, according to the latest data.* U.S. Department of Energy. https://www.eia.gov/pressroom/releases/press535.php

[6] Leskovec, J., Rajaraman, A., & Ullman, J. (2020). *Mining of Massive Datasets* (3rd ed.). Retrieved from http://www.mmds.org

[7] U.S. Energy Information Administration. (2023, December 18). *Energy use in homes.* Retrieved from https://www.eia.gov/energyexplained/use-of-energy/homes.php

[8] GeeksforGeeks. (2025, December 03). *Frequent Pattern Growth Algorithm.* Retrieved from https://www.geeksforgeeks.org/machine-learning/frequent-pattern-growth-algorithm/