CMPE- 255 Data Mining

The Ocho

Clustering Individual Household Electric Power Consumption

Presented to

Dr. Gheorghi Guzun

By

David Montes | SJSU ID: 015312115

Aaron Choi | SJSU ID: 015301117

William Carrasco | SJSU ID: 015288325

GitHub Repo: https://github.com/wilcarrasco/CMPE-255-Project

Contents

Introduction

Clustering Individual Household Electric Power Consumption

Motivation

Power usage and utilization is a metric that every household generates. Can patterns exist in time lapsed measurands and can that data be used to predict high demand or categorize usage? Patterns can be found in recurring daily tasks and making a large set to analyze these patterns can be beneficial to consumers and providers.

Objective

Our group proposes to use the Individual household electric power consumption data set to look for power consumption trends over time. We plan on clustering the data using descriptive methods to discover patterns and trends.

System Design & Implementation details

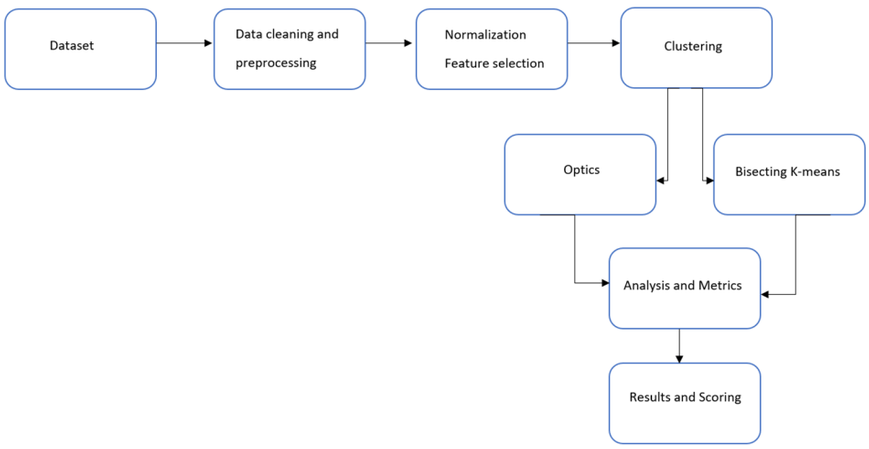
Algorithms considered/selected

- K-Means
- Bisecting K-Means
- PCA / SVD / TSNE
- L-2 / L-1 Normalization
- Feature selection / Extraction
- Optics w/ DBScan

Technologies and tools used

- Python 3.x
- Jupyter Notebooks
- VSCode, Numpy, Scipy, Sklearn, Matplotlib

System design/architecture/data flow



Experiments / Proof of concept evaluation

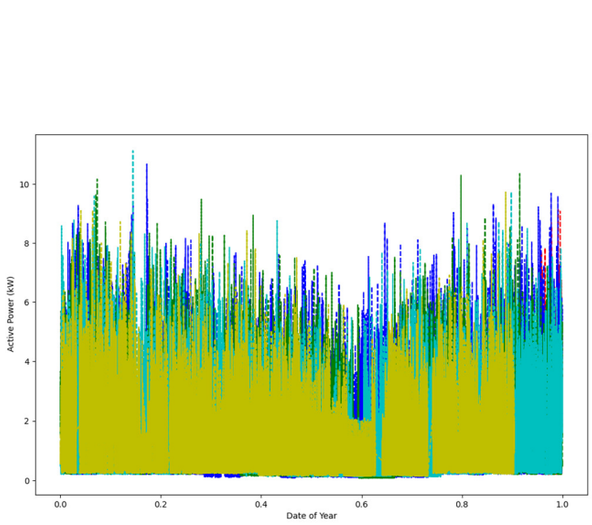
Dataset

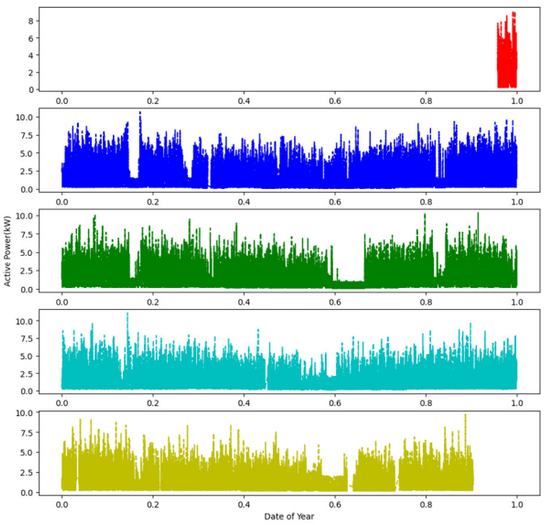
https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption

The dataset itself contains 2,075,259 power consumption measurements from a house in Sceaux, France between December 2006 and November 2010. The measurements were taken every minute and consist of the following:

- Date in dd/mm/yyyy
- Time in hh:mm:ss
- Global Active Power: household global minute-averaged power (kilowatt)
- Global Reactive Power: household minute-averaged reactive power (kilowatt)
- Voltage: minute-averaged voltage(volt)
- Global Intensity: household global minute-averaged current intensity (ampere)
- Sub Metering 1: Energy sub-metering which corresponds to the kitchen, containing a
- dishwasher, an oven, and a microwave. (watt-hour of active energy)
- Sub Metering 2: Energy sub-metering which corresponds to the laundry room, containing
- a washing-machine, a tumble-drier, a refrigerator, and a light
- Sub Metering 3: Energy sub-metering which corresponds to an electric water heater and
- an air-conditioner

Before choosing which method of dimensionality reduction to use, we took a look at the data. Initial observations were made to look at the global active power vs time at a year by year. When choosing the different dimensionality reduction techniques, we took a look at multiple techniques: PCA, SVD, t-SNE, and more.



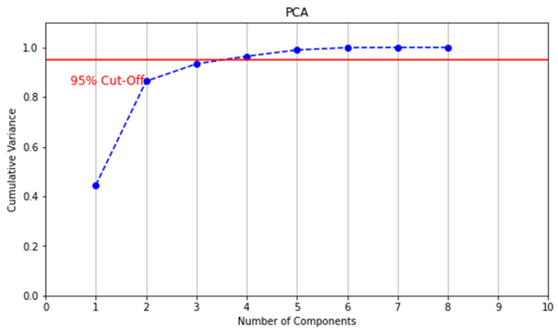


Preprocessing

The first step was to read the raw data and remove unwanted delimiter and next line characters. We opted to remove any uncaptured data from the original dataset as this lack of information would not be useful in clustering. This was done by removing the data with the '?' in columns 3 thru 9. Since the time data was in string format, we also converted that information into a numeric ratio and reduced, merging them into a single dimension instead of two. The remaining data was normalized by the max value in each column and checked the global power to check that the trends were the same. We applied multiple dimensionality reduction techniques such as PCA, SVD, and feature selection and determined that a combination of feature selection/extraction worked best.

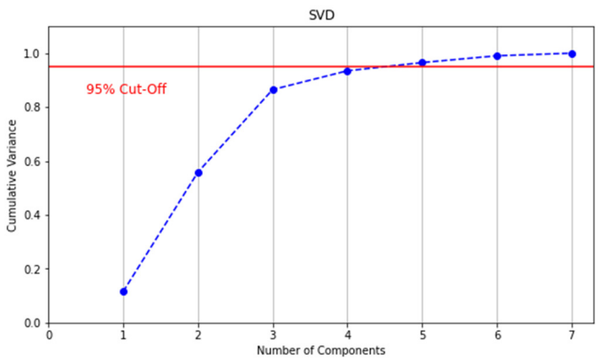
During the initial dimensionality reduction techniques, we took a look at PCA, SVD, and t-SNE. After the time conversion was made to the data, there were 7 columns containing power consumption data. The next step we approached was to normalize the data by the max value in the columns. After the normalization, we went to the next step with dimensionality reduction.

For PCA, we had taken an analysis of how many n_components we would use. Our threshold would be above 95% explained variance.



With 4 numbers of components, we could get a 96.44% explained variance.

For SVD, we had taken an analysis of how many n_components we would use. Our threshold would be above 95% explained variance.



With 5 numbers of components, we could get a 96.51% explained variance.

While observing the data even more, we noticed that submetering 1, 2, 3 may provide a better understanding of power consumption trends and relations. This would make sense since the global active power represents the active energy consumed by electrical equipment not measured in sub-metering 1, 2, and 3. When researching more about household power consumption, most of the electricity consumed appliances come from sub-metering 3, which includes electric water-heater and air-conditioning. The second most electrical power drawing equipment would be the sub-metering 2, which includes the washing-machine, tumble-drier, fridge, and lighting. Sub-metering.

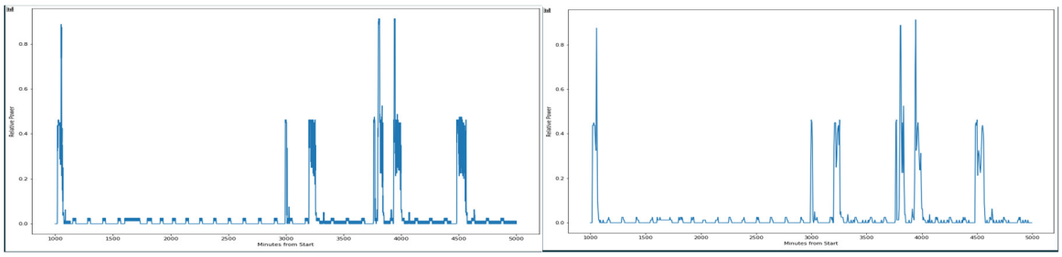
This can be confirmed by observing the original data of the sub-metering. Column 2, which represents submetering 3 (water-heater & AC), has the greatest mean compared to Column 1 (laundry room appliances) and Column 0 (kitchen appliances).

	0	1	2
count	2.049280e+06	2.049280e+06	2.049280e+06
mean	1.121923e+00	1.298520e+00	6.458447e+00
std	6.153031e+00	5.822026e+00	8.437154e+00
min	0.000000e+00	0.000000e+00	0.000000e+00
25%	0.000000e+00	0.000000e+00	0.000000e+00
50%	0.000000e+00	0.000000e+00	1.000000e+00
75%	0.000000e+00	1.000000e+00	1.700000e+01
max	8.800000e+01	8.000000e+01	3.100000e+01

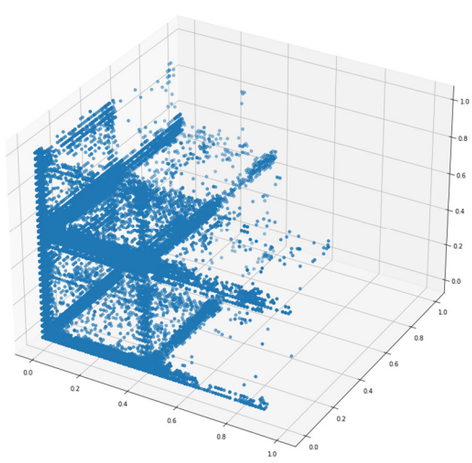
Therefore, we chose to choose only the 3 sub-metering data.

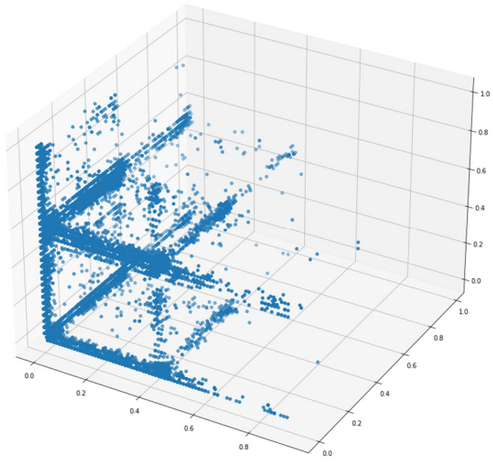
We also took the first two features (date and time) from the data and converted it into a single real value. This value was the number of minutes from the first time stamp (16/12/2006 17:24:00). The time would later be converted into day, week, and year ratios.

During our analysis, we found that working with 2 million data points was proving to be difficult without the aid of high-performance computing. However, considering the nature of this data helped with data reduction. We hypothesized that the majority power-consuming tasks in a household are not limited to a single minute. Therefore, it seemed reasonable to sample the data at larger intervals. The figures below show the relative power of sub-metering 2 from 1000 to 5000 minutes at 1 minute and 5-minute intervals, respectively. It is apparent that sampling at 1 minute only introduces noise and sampling at 5 minutes still captures the larger trends. Therefore, we chose to significantly reduce the data set by extracting every 5th sample.



Finally, to further improve processing time, we removed the zero-points (all three sub-metering's recorded 0) and used the test-train split function from Sci-Kit Learn and clustered 75% of the data. These data reduction techniques altogether resulted in 218,080 unique samples, or 10.5% of the original set. The difference between the raw and reduced dataset in 3 dimensions is shown in the figures below.

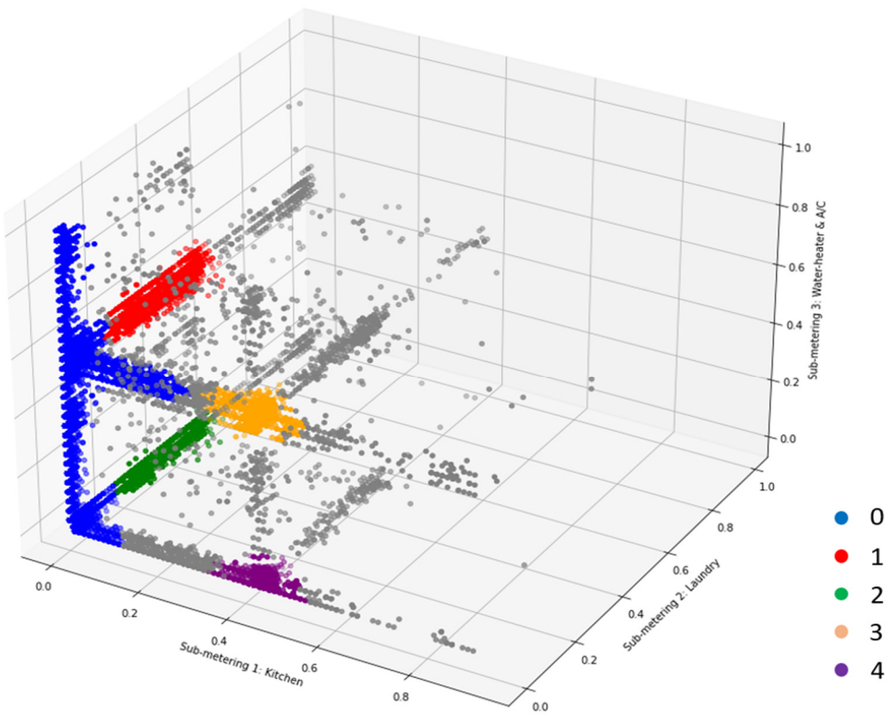




Methodology

We started by trying to cluster the samples with K-means, however, we found that the algorithm did not perform well with oblong groupings of the data. Considering the nature of the problem and the non-globular clusters, a density-based approach seemed more appropriate. Additionally, we wanted to exclude noisy points from the clusters, since our focus was on common usage samples. Therefore, DBSCAN was our algorithm of choice.

Graphs



Analysis

Since there was no ground truth data, we used the clustering sklearn.metrics library. To score the clustering, we used three metrics for the cluster: silhouette, Davies-Bouldin, and Calinski-Harabasz scores. The silhouette score helps similar the data to its own cluster. The Davies-Bouldin helps us observe the average similarity of each cluster to its most similar cluster. The Calinski Harabasz, also known as the Variance Ratio Criterion, is the score of ratios between within-cluster dispersion and between-cluster dispersion. For our cluster, we obtained the following results:

	OPTICS	KMeans
Silhouette Average	0.66	0.85
Davies-Bouldin	1.35	0.38
Calinski Harabasz	42314.28	674483.97

For the OPTICS clustering, we can see that our clustering had decent results. The silhouette score ranges from -1 to 1, where -1 would represent that the samples were assigned to the wrong cluster and 0 would indicate overlapping clusters. Our silhouette average of 0.6596 shows that data assigned to the clusters were relatively well assigned. The Davies-Bouldin result also shows that our clustering was apart and dispersed. For the Davies-Bouldin score, 0 would indicate the best clustering. Although our Davies-Bouldin score was not zero, we could observe some clusters were close. Our Calinski Harabasz score showed that the clustering was suitable since the clusters were not really spherical.

Although KMeans shows a stronger clustering, we also notice that it clusters all the data provided whereas OPTICS takes noise into consideration.

After the clustering with DBSCAN, we attempted to associate each cluster with a range of times. We wanted to show that a given cluster could correspond to a certain pattern of usage. For example, does a cluster with a high relative power output from sub-metering 2 (the laundry room) primarily contain timestamps from a certain day of the week? That might imply the resident of the house had a laundry day.

For this analysis, we extracted the timestamps from each cluster and day, week, and year ratio. The time ratios are decimal values from 0.0 to 1.0 which represent a fraction of the unit of time from the zero-point (in this case, the first timestamp in the data). The first time stamp is 16/12/2006 17:24:00, which is 17:24 (5:24 PM) Saturday, December 16th. Therefore, the zero-point is 17:24 for the day ratio, Saturday at 17:24 for the week ratio, and December 16th at 17:24 for the year ratio. This means a day ratio of 0.5 represents half a day from 17:24, which is 05:24. a week ratio of 1/7 is Monday at 17:24 and so on.

Once the time ratios were extracted for each cluster, we binned them into histograms for each cluster (shown below). For readability, the day histograms have 24 bins, the week histograms have 7, and the year histograms have 12.

The histograms which correspond to cluster 0 (blue) do not show any significant trends. This can likely be attributed to the fact that the cluster covers too wide of a scope. We ignore this cluster for the rest of our analysis.

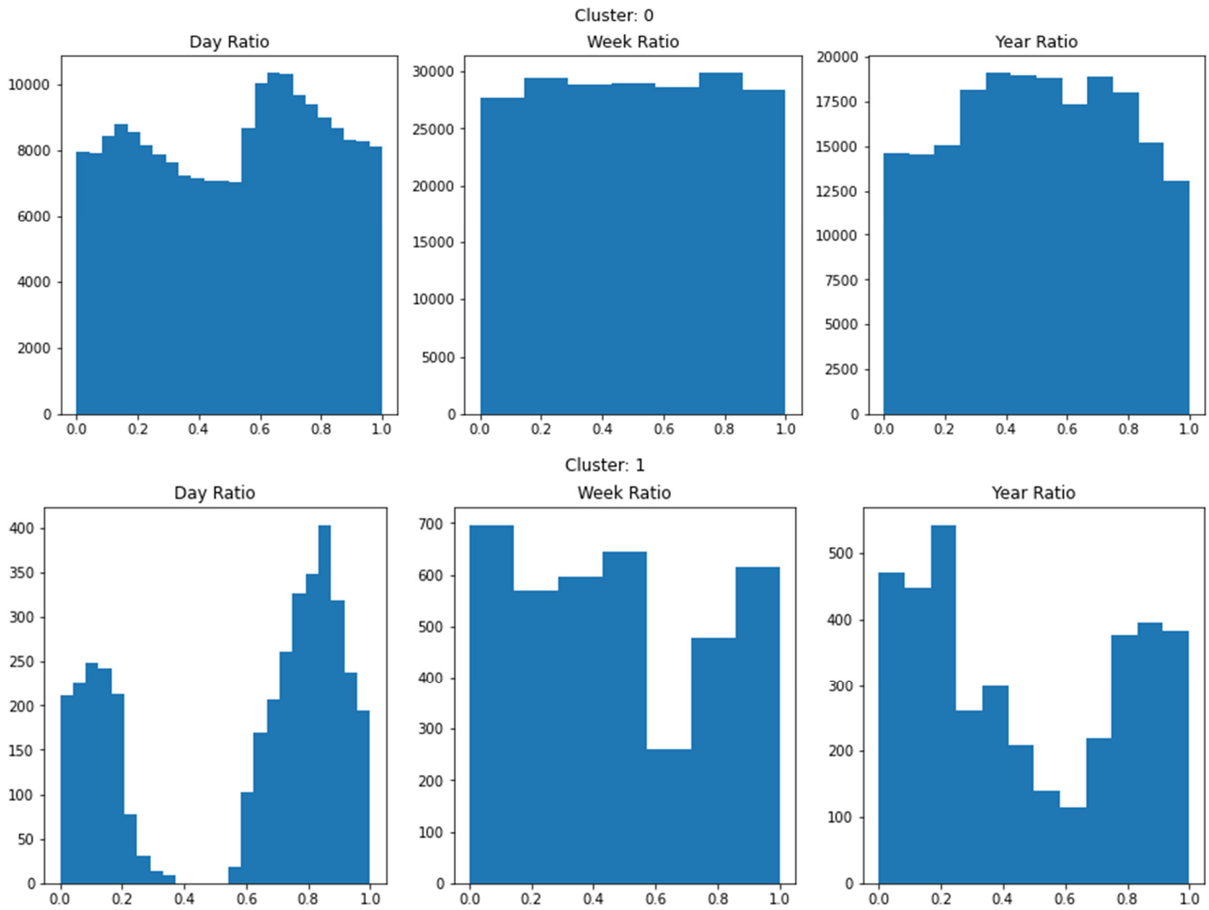
We identify the clusters based on a heuristic description below:

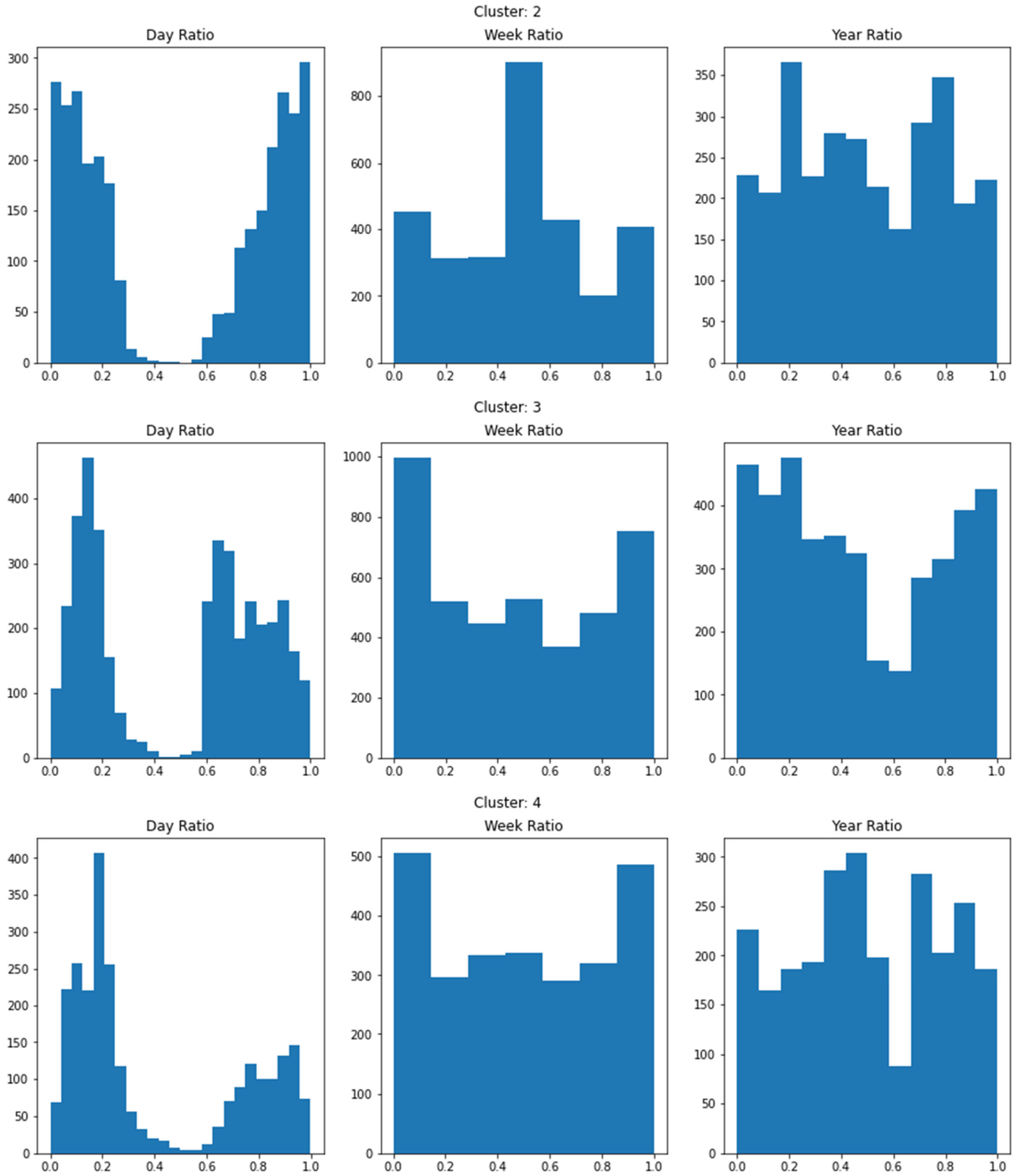
- Cluster 1 (Red): High Water-heater/AC usage with a high spread of laundry room usage
- Cluster 2 (Green): High spread of laundry room usage
- Cluster 3: (Orange): High Kitchen usage and high Water-heater/AC usage
- Cluster 4: (Purple): High Kitchen usage

First of all, every cluster showed a similar trend in the day ratio. In general, there is a dip in the number of samples from ~0.3 to ~0.7, which corresponds to ~01:00 and ~09:00. That's about 8 hours of sleep! In addition, there seems to be a relatively high number of high kitchen usage samples around 7:00 PM. This could be approximately the time the residents prepared dinner.

The week ratios also show some interesting trends. Both of the clusters with high Kitchen usage show spikes on Saturday and Sunday. This could be due to a tendency to use the kitchen throughout the day on weekends. During a weekday, more time is spent away from home. Also, there is a fascinating spike in cluster 2 for the middle of the week. Cluster 2 is the laundry room cluster, so it is possible that the residents of this home had a tendency to do their laundry in the middle of the week (possibly Wednesday or Tuesday).

Finally, the most significant trend in the yearly ratios is the dip in samples during the summer months for all clusters. This seems counter-intuitive since we typically associate higher rates of power consumption with warmer weather. However, it is possible that the residents of this house took yearly vacations.





Discussion & Conclusions

Decisions made

We decided against doing any regression task given the time constraints and not covering this topic in class. Initially we decided to use k-means as our clustering algorithm but had to choose another since the data was

not uniform in nature. We decided to use a feature selection technique that allowed us to better visualize the data and map it to our time ratio. Lastly, we wanted to preserve the time ratio metric but didn't want to normalize this data as some information would have been lost.

Difficulties faced

Our initial runs of DBSCAN from Sci-kit Learn resulted in memory errors. Upon further research, we found that the sklearn implementation is memory-inefficient due to its calculation of the full distance matrix for the samples. We then chose to use the OPTICS implementation, with the DBSCAN algorithm.

It was difficult to represent all data into a visual format down to two or three dimensions. There was a lot of noise and generally non-globular data to be able to use k-means. Dimensionality reduction techniques also failed to show patterns in the data as most more merged/overlapped amongst each other. Processing times for graphing such a large data set took a long time to complete. Additionally, the best clustering method, Optics w/ DBScan, took hours to compute.

Things that worked

Feature selecting/extraction on the three metering measurand yielded the best representation of the data and allowed for better clustering of usage metrics. The time ratio conversions also allowed us to represent the string time data into a quantifiable value that was used to compare meter metrics given certain time intervals. The Optics library with DBScan was critical in representing clusters since the non-globular shapes of the data. Lastly, modeling the data in 3D gave depth and more meaning to the cluster representations allowing for better analysis of the data.

Things that didn't work well

Since the dataset was quite large, we decided to apply PCA to reduce the dimensionality of the data. With the time ratio and massive amount of data it became difficult to sort through the data. Additionally, we tried truncated SVD with the same measure of success. TSNE was also tried but the space requirements for a successful run became unfeasible. TSNE is also computationally demanding, requiring a long time to compute. Since the dataset was not globular in nature, K-means had a hard time clustering groups successfully and generally merge swaths of clusters together.

Conclusion

This report has shown that interesting lifestyle trends can be derived from even a rudimentary clustering analysis of power consumption from only 3 separate rooms. From the results, we were able to posit that the residents of the house took summer vacations, did their laundry in the middle of the week, went to work on weekdays, and had a normal sleep schedule.

Project Plan / Task Distribution

Task	**Responsibility**
Dataset Selection	ALL
Data Exploration	ALL
Data Cleaning	David, Wil
Data Preprocessing	David, Wil
Time ration conversion	David
Dimensionality Reduction	Wil, Aaron

Graphs	All
Bisecting K-Means	Wil
Optics w/ DBScan	David, Aaron
Report	All
PowerPoint Presentation	TBD

GitHub Repo: https://github.com/wilcarrasco/CMPE-255-Project