# Watts the Matter?

## Classifying and Predicting Energy Consumption in Regions of the Eastern U.S.

YoungIn Kang, Ingrid Wijaya, Daniel Wray, Jessica Wong

# Introduction

In this project, we explored both unsupervised and supervised machine learning based on the features of twelve time series. The data came from PJM Hourly Energy Consumption Data on Kaggle, which included hourly energy consumption in megawatts (MW) for regions served by PJM Interconnection LLC (PJM) [12]. PJM provides electricity transmission in the Eastern United States, and each time series represented a different region served by PJM. Data from the Kaggle source was collected from the PJM website.

Supervised learning involves data where a set of features have been measured for a given number of observations, and a response variable has been measured for these observations as well. With methods for supervised learning, the objective is to predict the response variable using the features. Supervised learning involves both classification, in which the response variable is categorical, and regression, in which the response variable is quantitative. Some possibilities within classification include k-nearest neighbors, logistic regression, and discriminant analysis. Some possibilities within regression to strengthen models include subset selection, shrinkage, dimension reduction, and tree-based methods such as boosting, bagging, and random forests. In Section 3 of this project, we explored supervised learning by predicting future values of our time series using Facebook's Prophet.

On the other hand, unsupervised learning involves only data with the features measured for a given number of observations, without a response variable. Many methods for unsupervised learning involve identifying subgroups within the data, and it is often involved in exploratory data analysis. It can sometimes pose a challenge in that there is not as much of a specific objective as supervised learning. Some possibilities in unsupervised learning include principal component analysis, k-means clustering, and hierarchical clustering. In Section 2 of this project, we explored k-means clustering.

# 1 Data cleaning and data description

Each of the time series from Kaggle was in the form of a CSV file with observations recorded, an hourly frequency, with the first column providing the timestamp and the second column providing the energy consumption in megawatts (MW). The twelve time series, and their respective region names from the PJM website's zone map [15], were:

1. **AEP**: American Electric Power
2. **COMED**: Commonwealth Edison Company
3. **DAYTON**: Dayton Power and Light Co.
4. **DEOK**: Duke Energy Ohio and Kentucky Corp.
5. **DOM**: Dominion Energy
6. **DUQ**: Duquesne Light Co.
7. **EKPC**: East Kentucky Power Cooperative
8. **FE**: FirstEnergy Corp.
9. **NI**: Not available on the zone map, potentially represents PJM's N Illinois Hub [**pjm**]
10. **PJM Load**: Not available on the zone map, potentially represents aggregate PJM load from several regions prior to the start dates of the individual regions in this list
11. **PJME**: PJM/East. Not available on the zone map, likely represents the PJM East subregions according to the U.S. Energy Information Administration [19].
12. **PJMW**: PJM/West. Not available on the zone map, likely represents the PJM West subregion according to the U.S. Energy Information Administration [19].

Referring to the start and end dates in Table 1, the time series did not have a common time range between all twelve regions, so during this project we stored them in separate objects and made predictions for each time series individually.

Upon reading the CSVs into R, we sorted each time series by the timestamp. We then checked all time series for NA values and found none present. To clean the data, we checked each time series for duplicate entries, in which both the timestamp and the energy consumption figure matched another row. We did not find any, but we did find duplicate timestamps, as well as missing timestamps. Both the missing and duplicate timestamps corresponded to the beginnings and ends of Daylight Savings Time (DST). For duplicate timestamps, we removed the second row with the given timestamp. We did not find any timestamps that were associated with three or more observations.

The missing and duplicate timestamps we found correspond with the beginnings and ends of DST during the fall and spring of the years in Table 1 below. All of the DST beginnings that fell within the time range of each time series were duplicates, while only the DST endings from 2014-2017 were missing, so time series with end dates before 2014-2017 did not have any missing timestamps.

| Time Series (Frequency: Hourly, Source: PJM via Kaggle) | Start Timestamp | End Timestamp | Number of Observations After Removing Duplicates | Years with Missing Fall Timestamps (Start of DST) | Years with Duplicate Spring Timestamps (End of DST) |
|---|---|---|---|---|---|
| AEP | 10/1/2004, 1AM | 8/3/2018, 12AM | 121,269 | 2014-2017 | 2005-2018 |
| COMED | 1/1/2011, 1AM | 8/3/2018, 12AM | 66,493 | 2014-2017 | 2011-2018 |
| DAYTON | 10/1/2004, 1AM | 8/3/2018, 12AM | 121,271 | 2014-2017 | 2005-2018 |
| DEOK | 1/1/2012, 1AM | 8/3/2018, 12AM | 57,735 | 2014-2017 | 2012-2018 |
| DOM | 5/1/2005, 1AM | 8/3/2018, 12AM | 116,185 | 2014-2017 | 2006-2018 |
| DUQ | 1/1/2005, 1AM | 8/3/2018, 12AM | 119,064 | 2014-2017 | 2005-2018 |
| EKPC | 6/1/2013, 1AM | 8/3/2018, 12AM | 45,330 | 2014-2017 | 2014-2018 |
| FE | 6/1/2011, 1AM | 8/3/2018, 12AM | 62,870 | 2014-2017 | 2012-2018 |
| NI | 5/1/2004, 1AM | 1/1/2011, 12AM | 58,450 | None | 2005-2010 |
| PJM Load | 4/1/1998, 1AM | 1/1/2002, 12AM | 32,896 | None | 1998-2001 |
| PJME | 1/1/2002, 1AM | 8/3/2018, 12AM | 145,362 | 2014-2017 | 2002-2018 |
| PJMW | 4/1/2002, 1AM | 8/3/2018, 12AM | 143,202 | 2014-2017 | 2002-2018 |

Table 1: Start and end timestamps, number of observations after removing duplicates, missing timestamps, and duplicate timestamps for each time series

To find outliers, we plotted each time series and looked for any jumps in the energy consumption values. Because there were many observations, a series of relatively small or large observations could look like a jump but would appear smoother when we plotted a smaller window. To check for this possibility, we plotted the values around these potential outliers, and also inspected the raw values around them. After doing so, we identified four low outliers: one in DAYTON (Figure 1), one in DOM (Figure 2), one in FE (Figure 3), and one in PJMW (Figure 4). However, we did not remove or impute the values, this point and planned to check whether they would affect mean- or standard deviation-related features in the next section.

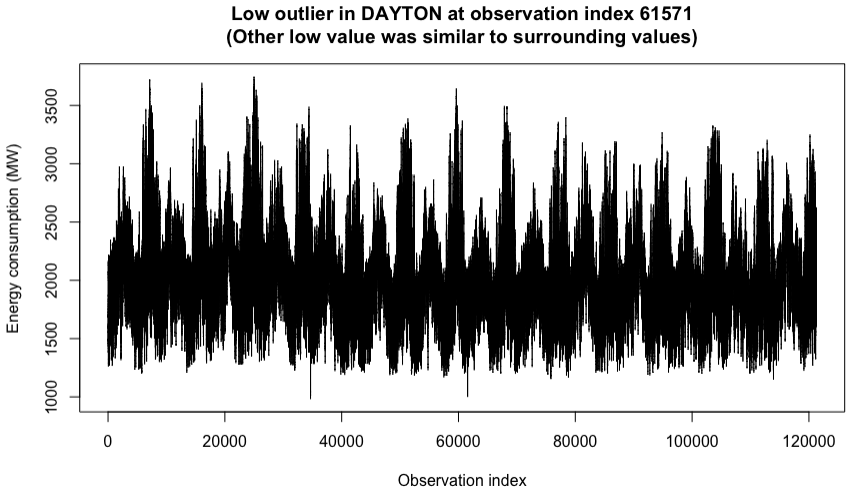


Figure 1: Outlier in DAYTON at index 61571

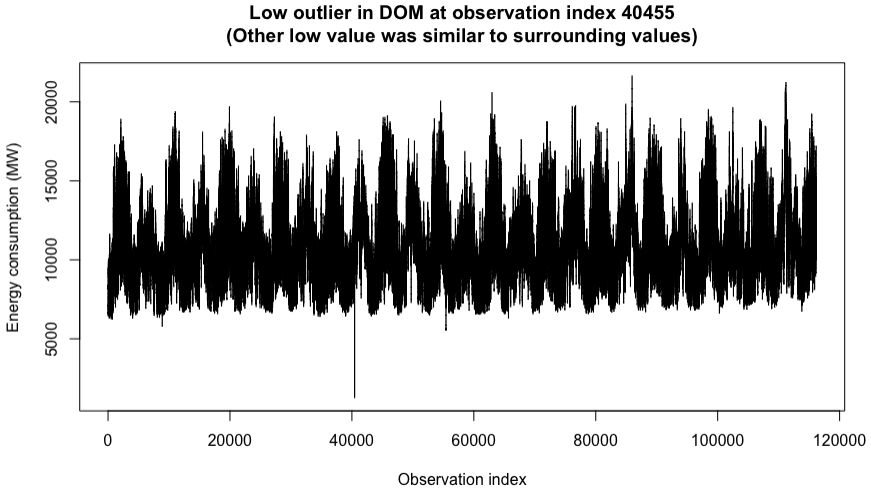


Figure 2: Outlier in DOM at index 40455

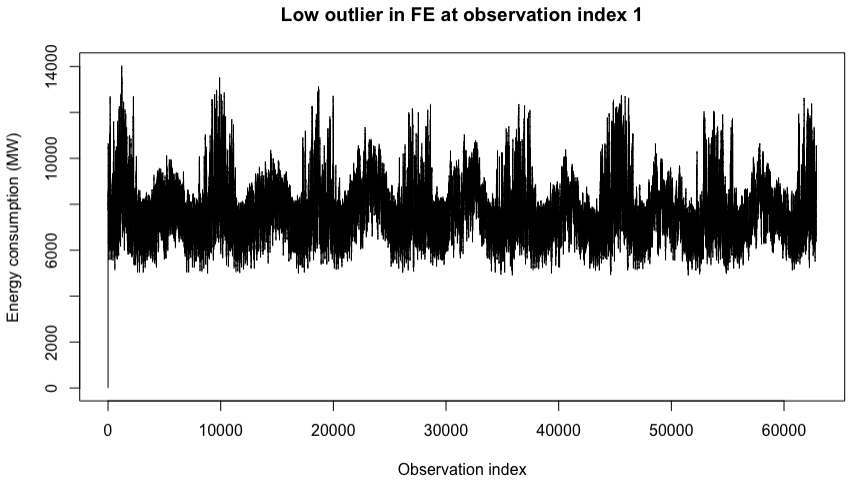


Figure 3: Outlier in FE at index 1

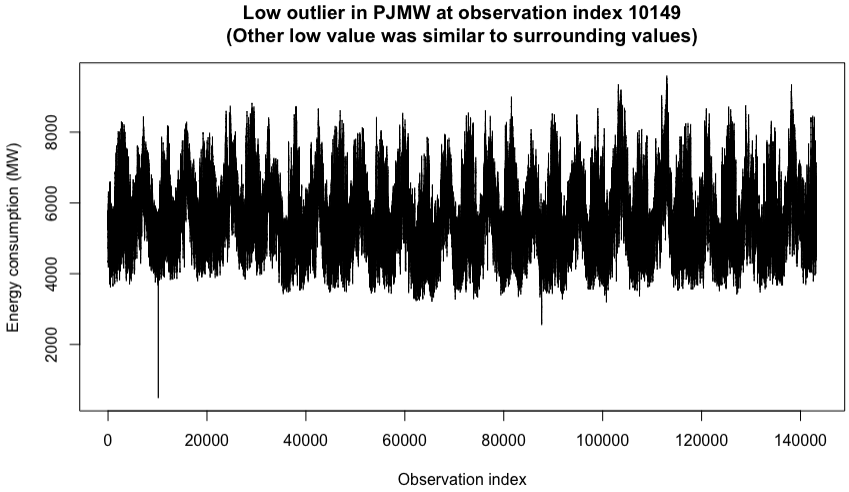


Figure 4: Outlier in PJMW at index 10149

Overall, we followed the same data cleaning process for each time series: We sorted by timestamp, checked for duplicate rows but did not find any, checked for duplicate timestamps, removed the second row for any duplicate timestamps, checked for missing timestamps, and checked for outliers. The differences lay in which time series had duplicate timestamps (all had four pairs of duplicates, except NI and PJM Load which had none) and which time series had outliers (DAYTON, DOM, FE, and PJMW, with one outlier each).

Once our data was cleaned, we proceeded to exploratory data analysis. For each time series, we created dygraphs to view different windows of the time series further, as well as seasonal box plots and ACF plots before any differencing to analyze seasonality at different cycles. We determined which seasonalities to consider for our cluster analysis based on the seasonalities that appeared significant based on these plots, and had similar findings across the time series. Below are two examples of our plots analysis to illustrate this process, using AEP and PJME.

## 1.1 EDA Example with Plots: AEP

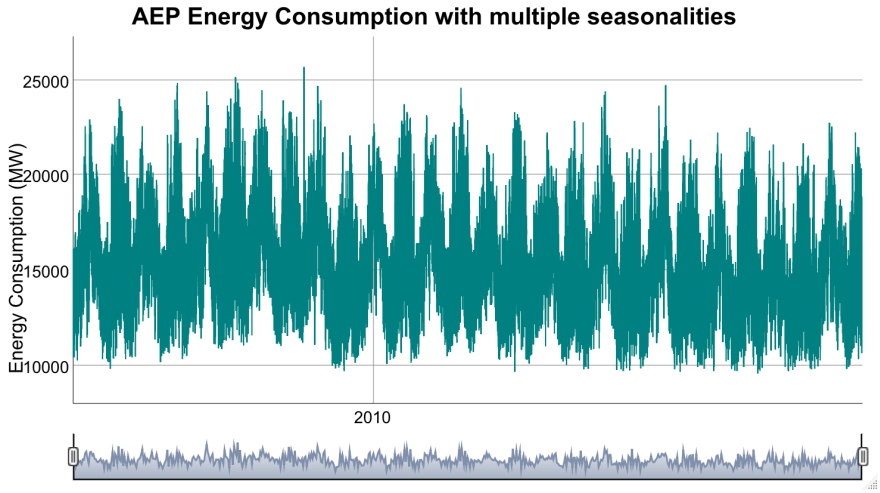


Figure 5: Dygraph of AEP

Referring to Figure 5, we did not see an overall trend of increasing or decreasing throughout the entire AEP time series, but we saw seasonality at various levels. We investigated the multiple seasonalities further through the seasonal box plots and ACF plots.

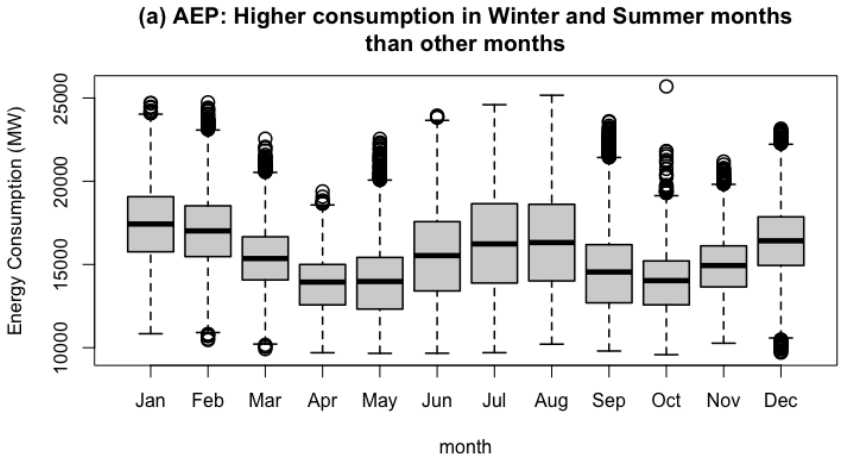


Figure 6: Seasonal boxplot of AEP by month

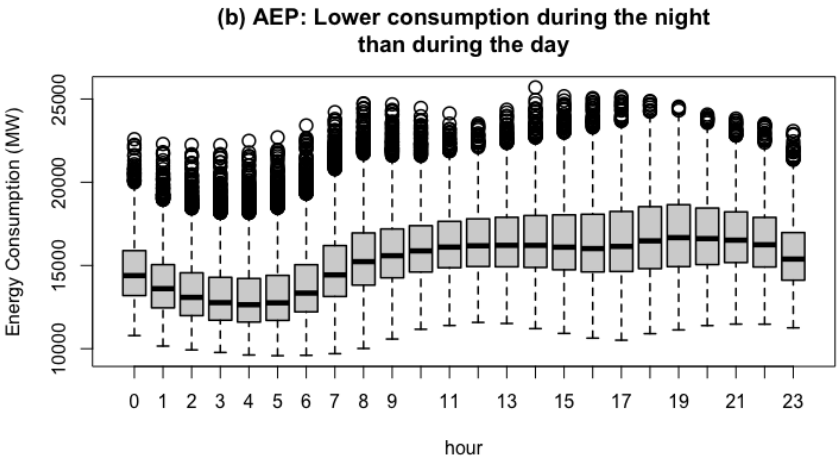


Figure 7: Seasonal boxplot of AEP by hour of the day

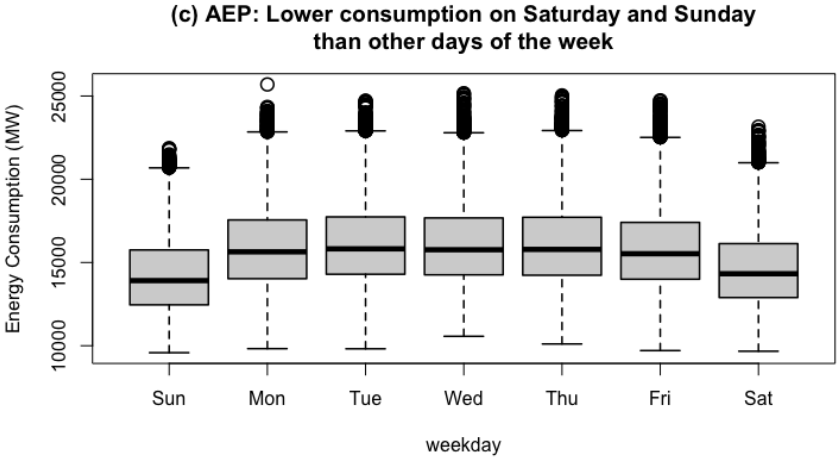


Figure 8: Seasonal boxplot of AEP by weekday

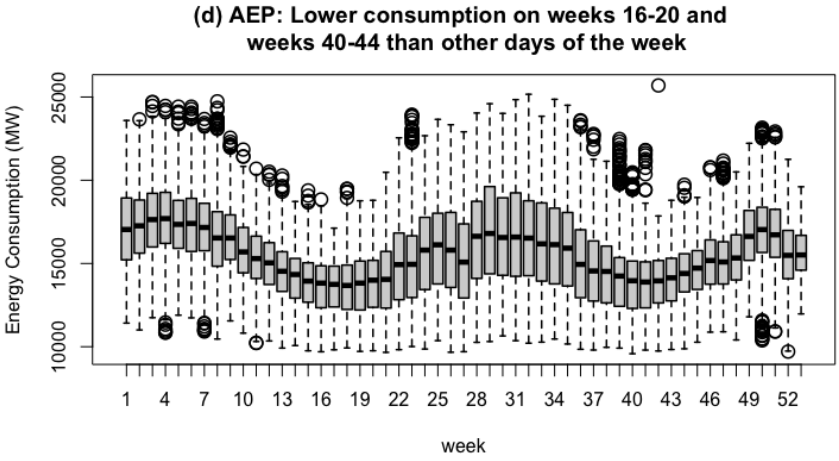


Figure 9: Seasonal boxplot of AEP by week of the year

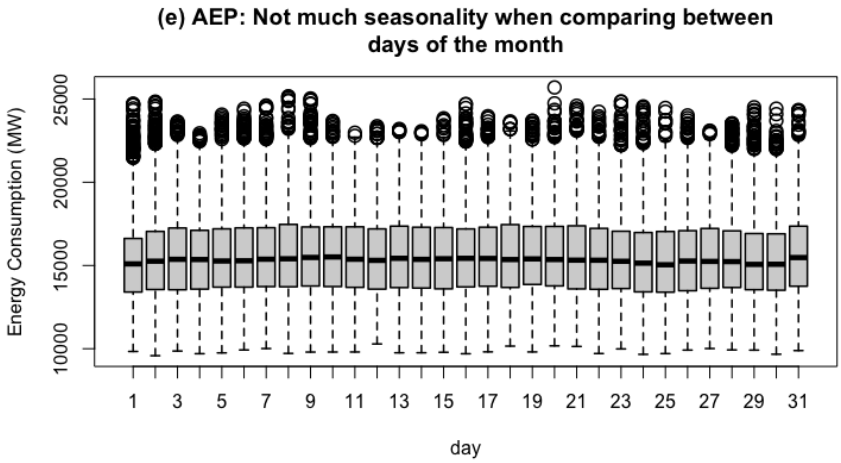


Figure 10: Seasonal boxplot of AEP by day of the month

Referring to Figures 6, 7, 8, and 9, we saw patterns of seasonality by month, hour of the day, weekday, and week of the year. Referring to Figure 10, we did not see much seasonality between days of the month. This suggested that we should focus on seasonality by month of the year, hour of the day, day of the week, and week of the year when performing our clustering, with month and week of the year showing similar information due to the same yearly cycle.

After the seasonal box plots, we next examined the ACF plots of AEP using different cycles to note patterns from multiple seasonalities.

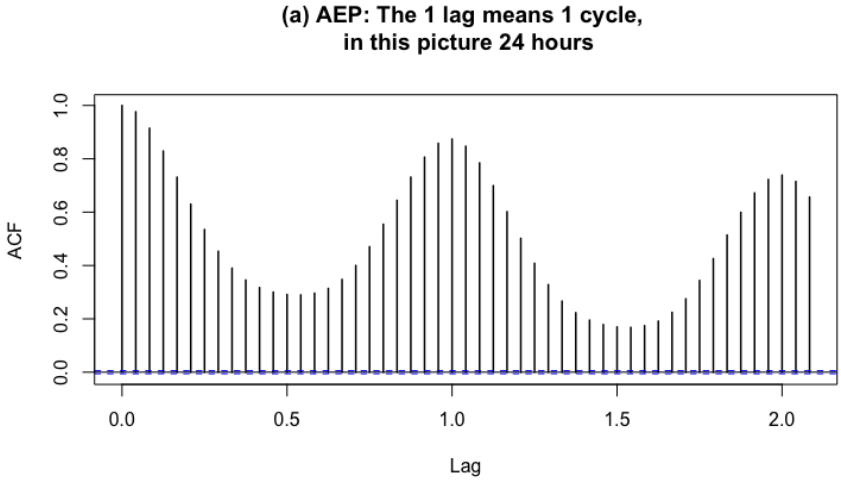


Figure 11: ACF plot of AEP using daily cycle

Referring to Figure 11, we noticed a pattern of gradual decreases followed by increases that reoccurred every 24 hours, or one day, confirming that daily seasonality would be useful to account for in our features.

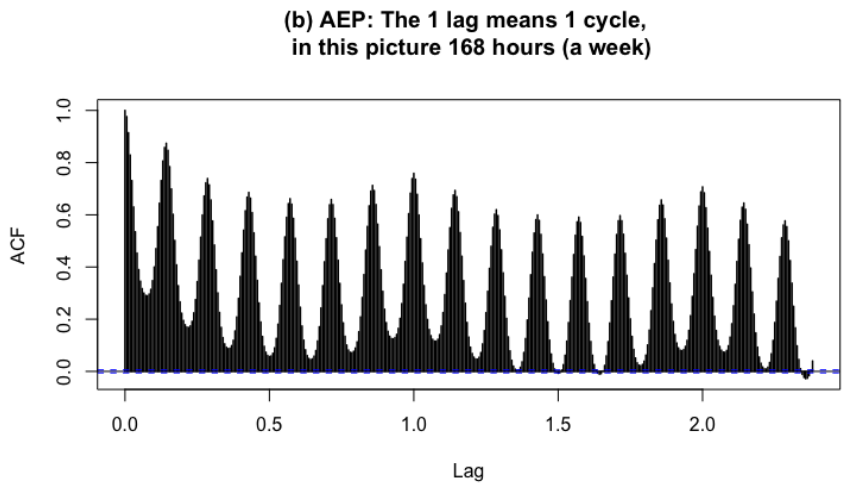


Figure 12: ACF plot of AEP using weekly cycle

Referring to figure 12, we can see the daily seasonality as well, but also an underlying pattern of gradual decreases and then increases that reoccurred every 168 hours, or 1 week, confirming that weekly seasonality would be useful to account for in our features.

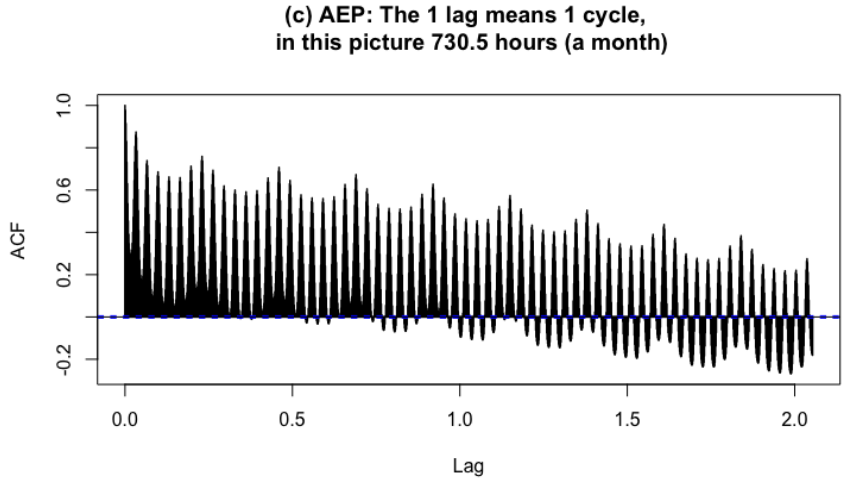


Figure 13: ACF plot of AEP using monthly cycle

Referring to figure 13, we can see the daily and weekly seasonality, as well as a gradual decreasing throughout the ACF plot. With the interference of these two other types of seasonality, there does not appear to be a cyclical pattern reoccurring approximately every 730.5 hours, so the day or week of the month may not be as important in terms of seasonality as the hour of the day or day of the week. However, we found the decreasing pattern to be interesting, so we decided to further investigate the monthly cycle in our features.

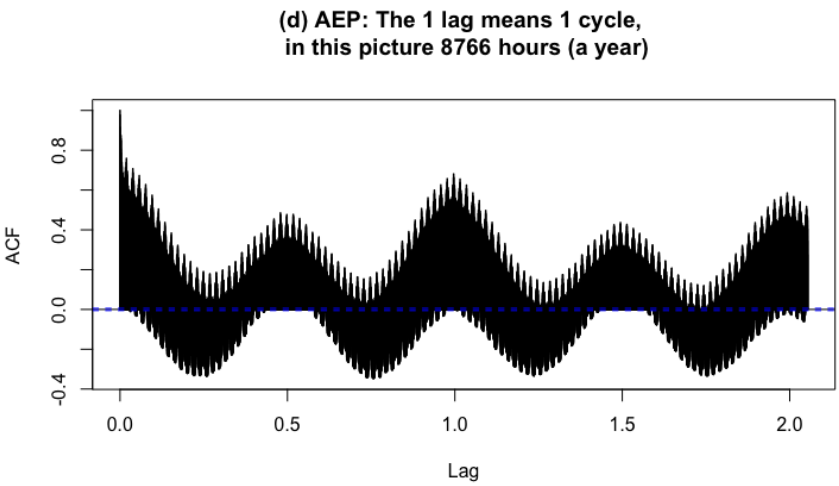


Figure 14: ACF plot of AEP using yearly cycle

Referring to figure 14, we can see the daily and weekly seasonality, as well as an underlying pattern of decreasing, increasing, decreasing, and then increasing again each year. This underlying pattern can be further investigated by looking at monthly seasonality to account for patterns repeating each yearly cycle. Seasonality by either the month or the week of the year could potentially relate to this pattern, as we saw in our seasonal box plots as well.

## 1.2 EDA Example with Plots: PJME

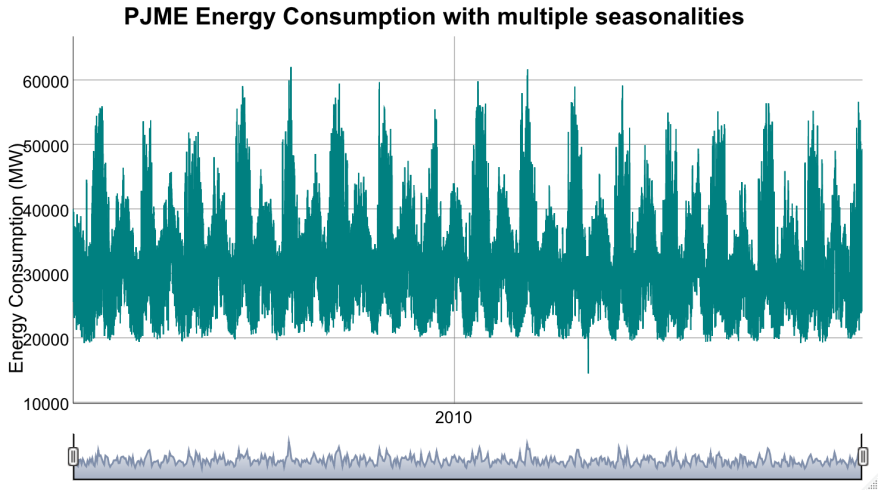


Figure 15: Dygraph of PJME

Referring to Figure 15, we did not see an overall trend of increasing or decreasing throughout the entire PJME time series, but we saw seasonality at various levels. We investigated the multiple seasonalities further through the seasonal box plots and ACF plots.

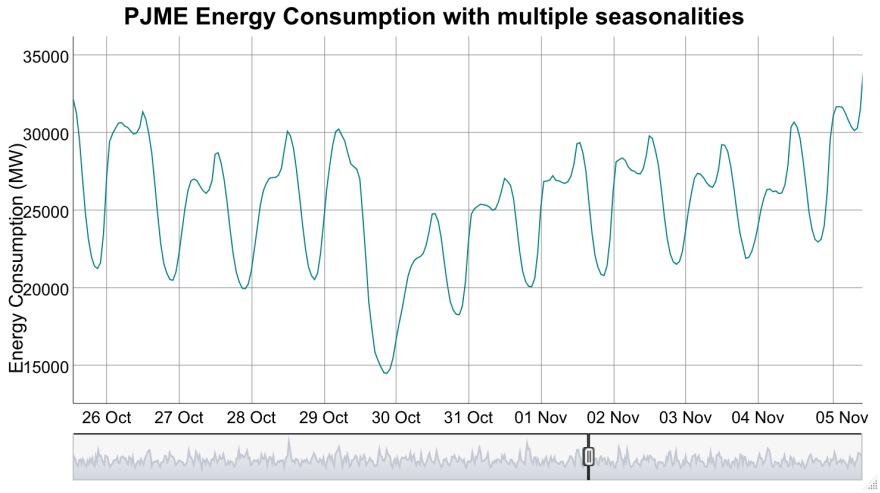


Figure 16: Dygraph of PJME zoomed in on low value

Referring to Figure 16, we used the dygraph to double check our finding from the the data cleaning that the low value visible in Figure 15 was not an outlier, and confirmed that it was close enough to its surrounding values to not be an outlier.

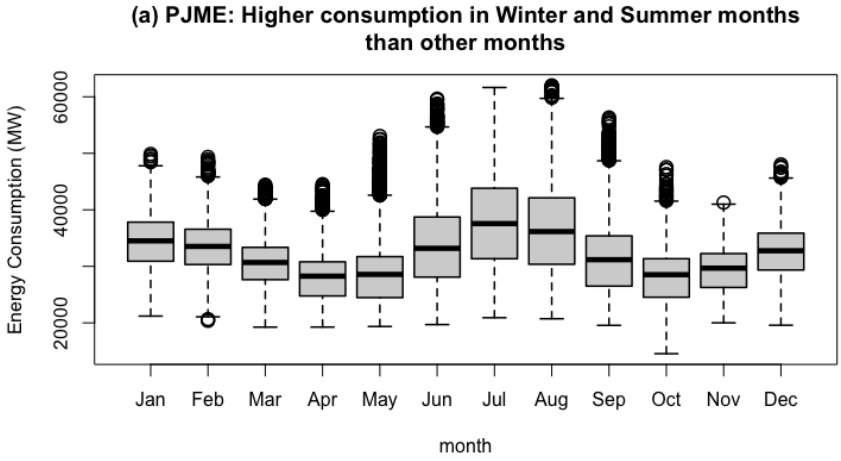


Figure 17: Seasonal boxplot of PJME by month

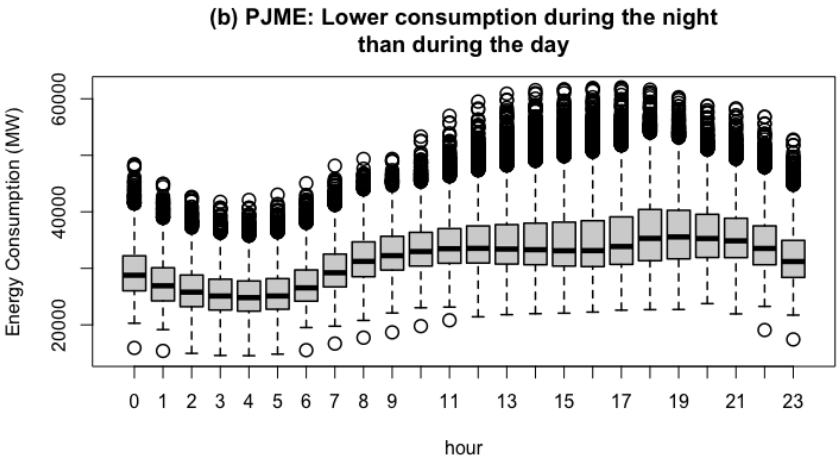


Figure 18: Seasonal boxplot of PJME by hour of the day

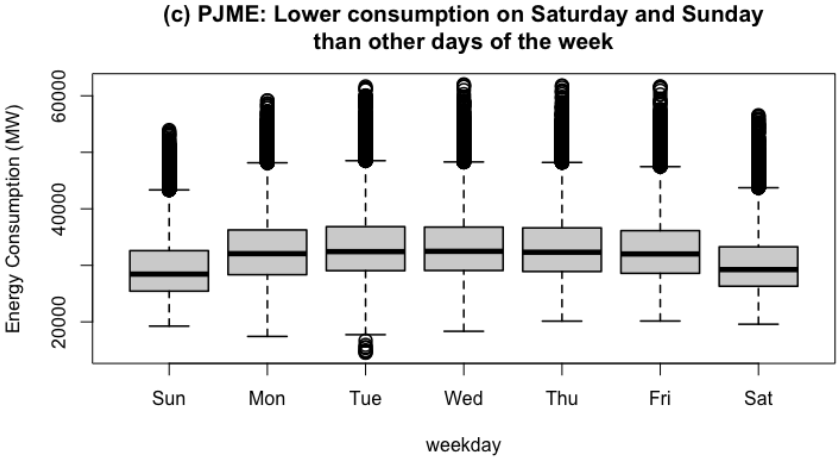


Figure 19: Seasonal boxplot of PJME by weekday

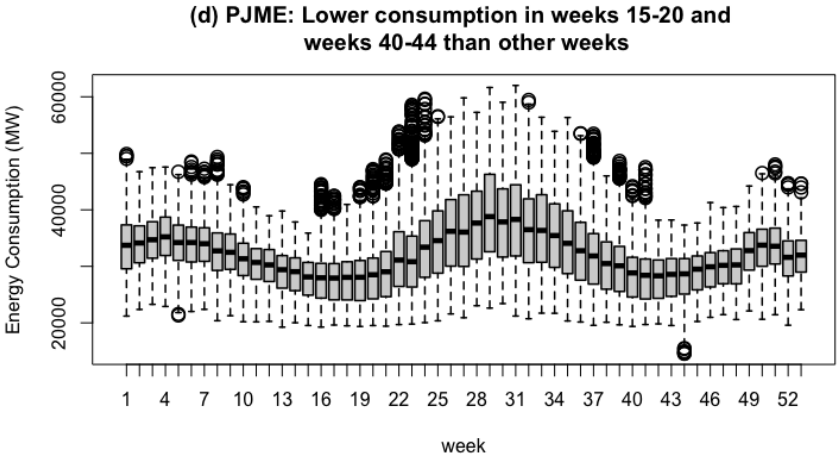


Figure 20: Seasonal boxplot of PJME by week of the year

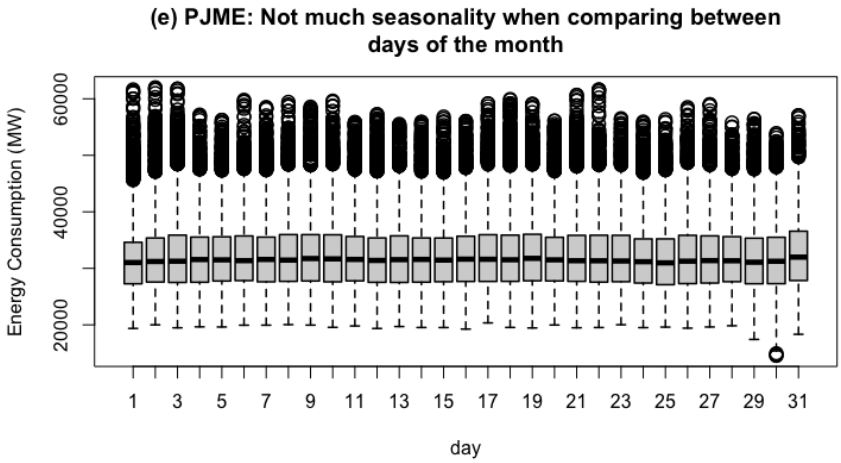


Figure 21: Seasonal boxplot of PJME by day of the month

Referring to Figures 17, 18, 19, and 20, we saw patterns of seasonality by month, hour of the day, weekday, and week of the year. Referring to Figure 21, we did not see much seasonality between days of the month. This suggested that we should focus on seasonality by month of the year, hour of the day, day of the week, and week of the year when performing our clustering, with month and week of the year showing similar information due to the same yearly cycle.

After the seasonal box plots, we next examined the ACF plots of PJME using different cycles to note patterns from multiple seasonalities.

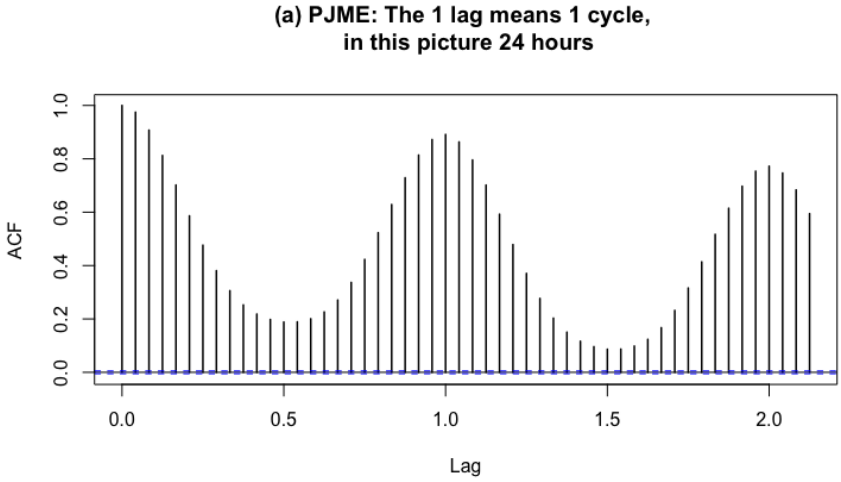


Figure 22: ACF plot of PJME using daily cycle

Referring to Figure 22, we noticed a pattern of gradual decreases followed by increases that reoccurred every 24 hours, or one day, confirming that daily seasonality would be useful to account for in our features.

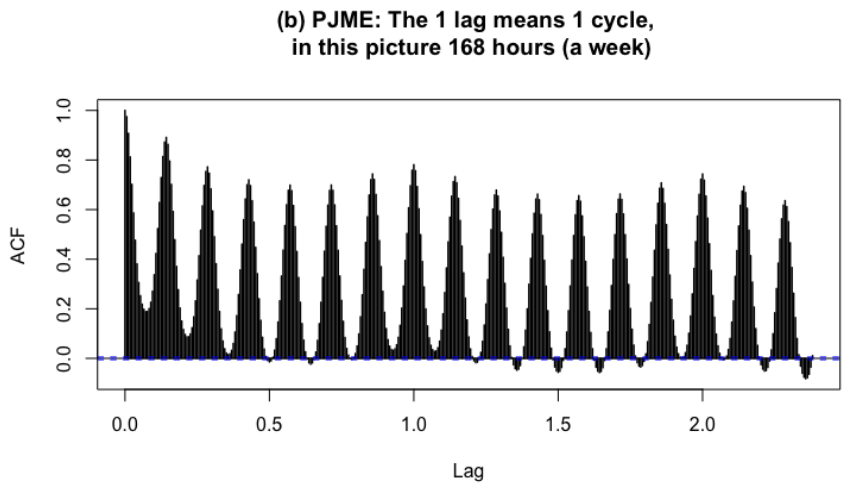


Figure 23: ACF plot of PJME using weekly cycle

Referring to figure 23, we can see the daily seasonality as well, but also an underlying pattern of gradual decreases and then increases that reoccurred every 168 hours, or 1 week, confirming that weekly seasonality would be useful to account for in our features.

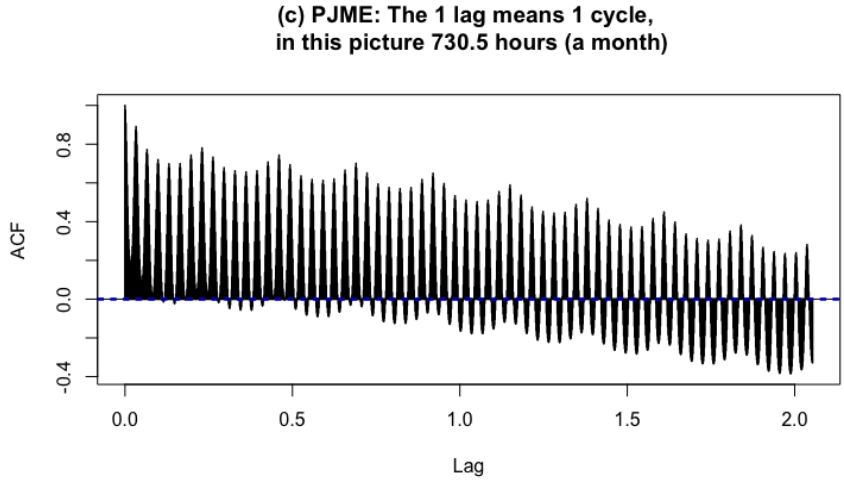


Figure 24: ACF plot of PJME using monthly cycle

Referring to figure 24, we can see the daily and weekly seasonality, as well as a gradual decreasing throughout the ACF plot. With the interference of these two other types of seasonality, there does not appear to be a cyclical pattern reoccurring approximately every 730.5 hours, so the day or week of the month may not be as important in terms of seasonality as the hour of the day or day of the week. However, we found the decreasing pattern to be interesting, so we decided to further investigate the monthly cycle in our features.

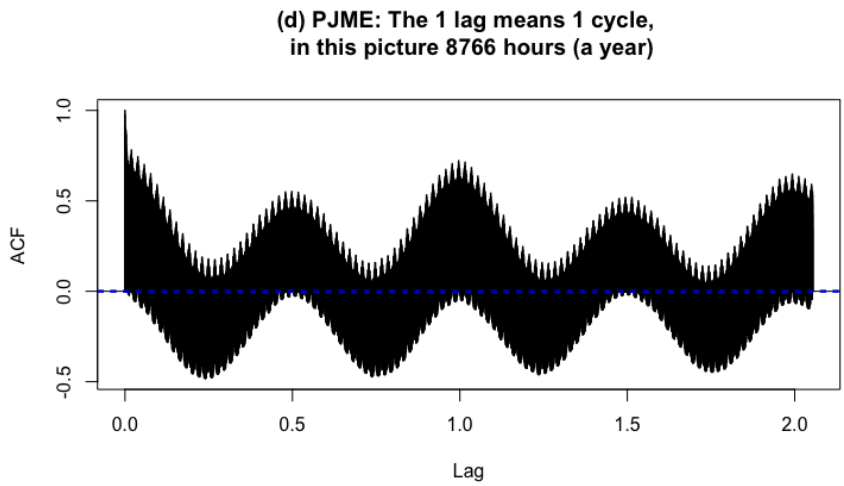


Figure 25: ACF plot of PJME using yearly cycle

Referring to figure 25, we can see the daily and weekly seasonality, as well as an underlying pattern of decreasing, increasing, decreasing, and then increasing again each year. This underlying pattern can be further investigated by looking at monthly seasonality to account for patterns repeating each yearly cycle. Seasonality by either the month or the week of the year could potentially relate to this pattern, as we saw in our seasonal box plots as well.

Based on the dygraphs, box plots, and ACF plots for the twelve time series, we confirmed that there were consistent findings among these time series: Seasonality by hour of the day, day of the week, week of the year, and month of the year (with week of the year and month of the year showing similar information due to the same yearly cycle) would be the most useful types of seasonality to explore with our features. Seasonality by day of the month was perhaps not as meaningful for finding seasonal patterns, but we did find an interesting decreasing pattern in the ACF plots. We thus included monthly cycles in our initial set of ACF- and PACF-related features in order to further investigate whether this pattern was mostly influenced by interference from other seasonalities or could provide a useful feature on its own.

We used these findings when creating features for clustering in the following section. Along with calculating summary statistics for each time series as a whole, we also calculated summary statistics among the averages we computed by each of the time increments we found useful when analyzing the multiple seasonalities with box plots and ACF plots: averages by hour, day of the week, week of the year, and month of the year. We also extracted the autocorrelations and partial autocorrelations around each of the the seasonal lags at daily, weekly, monthly, and yearly cycles, based on our findings from the ACF plots with different seasonalities.

# 2 Unsupervised machine learning

## 2.1 Feature Calculation

To start with, we created a total of 90 features. Of these, 48 were related to ACFs and PACFs, and were the first three regular and first three seasonal lags of the ACFs and PACFs for the time series when their frequency was made daily, weekly, monthly, and yearly.

Another 30 of our features were related to summary statistics: the mean, median, interquartile range, minimum, maximum, and standard deviation. These were calculated for the whole time series, for the average by month of the year, for the average by week of the year, for the average by day of the week, and for the average by hour of the day. Seasonality by day of the month did not seem significant using the seasonal box plots, so we did not calculate the features for this average.

We had not removed or imputed values for any outliers during our data cleaning process and planned to see whether the outliers would significantly affect the mean and standard deviation summary statistics, for both the overall time series and the averages by different time increments. We calculated the summary statistics with the outliers retained as is, with the outliers removed, and with the outliers imputed using the average value from the same days of the week that month (based on the suggestion in the discussion forum). We checked the changes in all summary statistic features to learn more about the outliers' effects, but for our decision in handling outliers we focused on the mean and standard deviation ones.

For both removing and imputing values for the outliers, we found that the mean and standard deviation summary statistics did not change by more than 5%. In fact, the changes were all under 1%, and often under 0.001%. We therefore decided not to remove or impute values for the outliers. However, we saw that keeping the outliers in would drastically affect the time series' overall minimum feature, so we decided to set the overall minimum for the affected time series (DOM, FE, and PJMW – DAYTON also had an outlier but it was not the overall minimum) to the second smallest value in the time series, since the values we would have imputed for the outliers were still greater than these second smallest values.

Our other features included proportion and number of observations above the mean, proportion and number of observations above 2 standard deviations, proportion of increases and decreases, slope of a line fitted to the time series, number of missing observations, optimal parameter for Box-Cox transformations, degree of differencing needed to remove seasonality and achieve stationarity, and sample size. A summary of our feature categories is in Table 2 below:

| Feature Category | Number of Features | Description |
|---|---|---|
| Autocorrelations | 24 | 3 regular lags and 3 seasonal lags for each daily, weekly, monthly and yearly cycle |
| Partial autocorrelations | 24 | 3 regular lags and 3 seasonal lags for each daily, weekly, monthly and yearly cycle |
| Missing observations | 1 | Number of missing data points in time series |
| Slope of linear model | 1 | Slope of linear model to quantify trend |
| Box-Cox transformation | 1 | Optimal parameter for Box-Cox transformation to stabilize variance |
| Trend | 2 | Proportion of positive and negative trend from each value to the next |
| Summary statistics | 30 | Mean, median, IQR, min, max, and SD of time series by hour, day of week, week, month, and as a whole |
| Order of differencing | 2 | Order of regular (d) and seasonal differencing (D) to observe improvements in stationarity |
| Sample size | 1 | Number of data points in time series |
| Above the mean | 2 | Percentage and number of data points that are above the mean |
| More than two standard deviations from mean | 2 | Percentage and number of data points that are more than two standard deviations away from mean, in either direction |
| **Total** | 90 | |

Table 2: Description of each feature category

## 2.2 K-Means Clustering

We first attempted K-means clustering to split the time series into 2 groups. Cluster analysis using K means obtained the following 2 clusters in Table 3:

| | |
|---|---|
| Cluster 1 | PJM Load, PJME |
| Cluster 2 | AEP, COMED, DAYTON, DEOK, DOM, DUQ, EKPC, FE, NE, PJMW |

Table 3: Clusters obtained from K-means with 2 clusters

Using percentage increase of cluster 2's mean from cluster 1's mean as an indicator of features with distinct means and 50% as our cut off point of a large difference between means. We obtained the following 40 features that have the most distinct means between the clusters:

- Autocorrelations of a yearly cycle at all the seasonal lags 1, 2, 3.
- Partial autocorrelations of a monthly cycle at all the seasonal lags 1, 2, 3.
- Partial autocorrelations of a yearly cycle at the seasonal lags 2, 3.
- Slope of the linear model.
- Optimal parameter for box-cox transformation
- Summary Statistics (Mean, Median, IQR, Min, Max,and Standard Deviation) of the time series
- Summary Statistics for the average by months of the year
- Summary Statistics for the average by weeks of the year
- Summary Statistics for the average by days of the week
- Summary Statistics for the average by hours of the day

Below in Table 4 are the means of the clusters for the top 10 features with the greatest absolute values in their percent changes between cluster means:

| Feature | Percent Change from Cluster 1 Mean to Cluster 2 Mean | Cluster 1 Mean | Cluster 2 Mean |
|---|---|---|---|
| ACF one lag after the yearly seasonal lag | 1663.916% | -0.002707543 | 0.042343698 |
| PACF at the monthly seasonal lag | 413.127% | -0.0009724277 | 0.0030449324 |
| PACF at the yearly seasonal lag | 177.361% | 0.001680581 | -0.001300110 |
| PACF one lag after the monthly seasonal lag | 141.220% | -0.0015908920 | 0.0006557702 |
| ACF at the yearly seasonal lag | -126.602% | 0.03945416 | 0.08940392 |
| Slope of linear model fit | 109.840% | 0.020817064 | -0.002048372 |
| Order of Box-Cox transformation | -102.500% | -0.1616162 | -0.3272727 |
| Standard deviation of the averages by hour of the day | 80.536% | 3858.3961 | 750.9997 |
| Interquartile range of the averages by hour of the day | 79.827% | 5969.059 | 1204.150 |
| Interquartile range of the averages by month of the year | 78.882% | 4601.5123 | 971.7586 |

Table 4: Approximate population served in each PJM region today from Cluster 1

Since we have 40 features that have distinct means between clusters, we selected several pairs of features that distinguished the clusters well in the pairwise scatter plots below (See Figure 26).

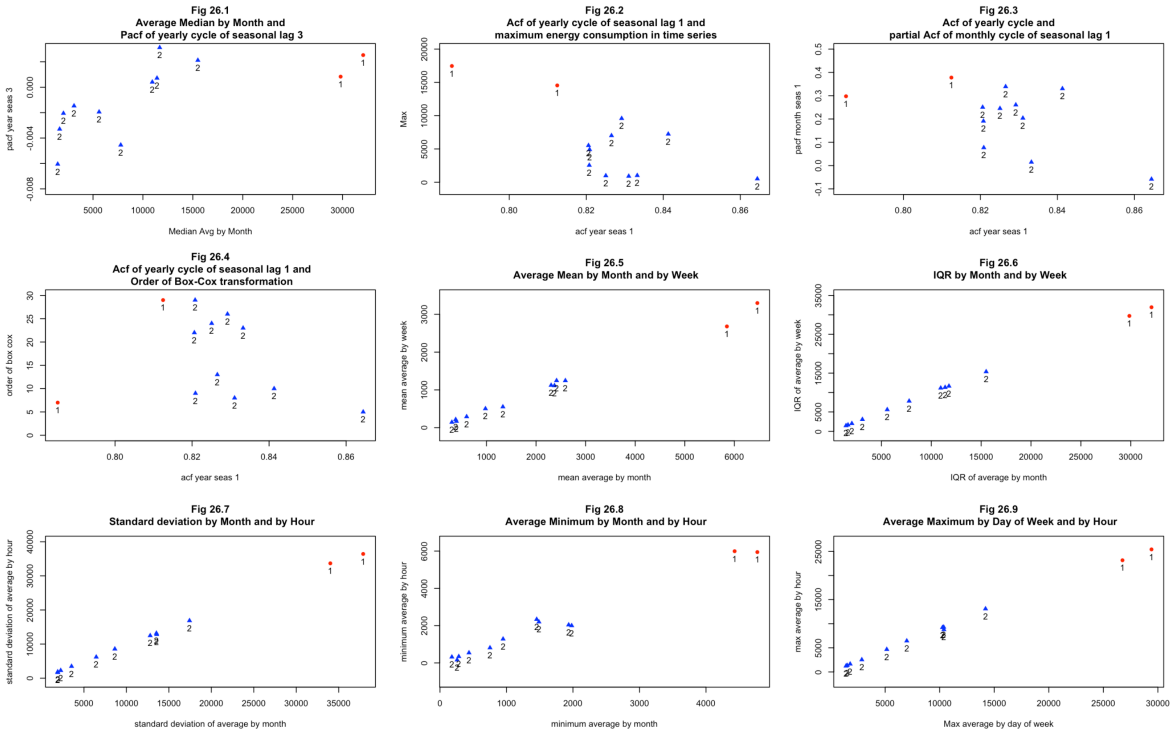


Figure 26: Pairwise scatter plots of features with distinct means after clustering into 2 groups

Referring to Fig 26.1 to 26.4 in Figure 26, it appears that amongst the features of autocorrelations and partial autocorrelations, the autocorrelatons of a yearly cycle at seasonal lag 1 stands out in terms of separating the clusters into 2. More specifically, Cluster 1 and Cluster 2 have mean autocorrelations at seasonal lag of 0.099 and 0.151 respectively.

Referring to Fig 26.1, 26.2, 26.5, 26.8 and 26.9 in Figure 26, we can observe that the average energy consumption by month, weeks, day of the week and hours of the day is significantly higher in cluster 1 as compared to cluster 2. For instance, in Fig 26.5, the average mean energy consumption by month and by weeks for cluster 1 is approximately from 5500MW to 6500MW, and 2500MW to 3500MW respectively. On the other hand, that for cluster 2 is approximately from 500MW to 2500MW, and 0MW to 1500MW respectively. Hence, we can conclude that energy consumption throughout the year is generally higher in cluster 1 as compared to cluster 2. This is probably because regions in Cluster 1 serves a larger population than regions in Cluster 2, which we will further analyse by looking at the customers of the company in each region.

According to U.S. Energy Information Administration and PJM Interconnection, Companies that are in charge of energy distribution in PJM/East (PJME) are Delmarva Power and Light, Pepco, Baltimore Gas and Electric, Atlantic City Electric, Peco Energy, Jersey Central Power and Light, PSE&G, Rockland Electric, Med-Ed, PPL Electric Utilities and Penelec. Additionally, Companies that are in charge of energy distribution in PJM/West(PJMW) are American Transmission Systems, American Electric Power, Dayton Power and Light Energy, Duke Energy Ohio Kentucky,Duquesne Light and Allegheny Power Systems [15][19]. We provided a table estimating the customers served in each region of each cluster below (See Table 5 and Table 6).

| Cluster 1 | Customers served (in millions) | Source |
|---|---|---|
| PJM Load | NA | NA |
| PJME | 11.096 | Companies under PJM/East [14, 3, 4, 6, 16, 11, 18, 13, 17] |

Table 5: Approximate population served in each PJM region today from Cluster 1

| Cluster 2 | Customers served (in millions) | Source |
|---|---|---|
| AEP | 5.5 | American Electric Power [1] |
| COMED | 4 | Commonwealth Edison Company [5] |
| DAYTON | 0.527 | Dayton Power and Light Company [2] |
| DEOK | 0.870 | Duke Energy [8] |
| DOM | 7 | Dominion Energy [7] |
| DUQ | 0.6 | Duquesne Light Company [9] |
| EKPC | 1.1 | East Kentucky Power Cooperative [10] |
| FE | 6 | First Energy Corporation [11] |
| NI | 4 | Commonwealth Edison Company [5] |
| PJMW | 9 | Companies under PJM/East [1, 2, 8, 9, 11] |

Table 6: Approximate population served in each PJM region today from Cluster 2

Based on the results of cluster analysis using K-means, we claimed that regions in Cluster 1 serves a larger population than regions in Cluster 2 is a plausible reason for the higher energy consumption in Cluster 1. According to Table 6, we can observe that indeed customers served in cluster 2 are generally smaller than in cluster 1. Although there is not much information on the customer served in PJM Load, recall that we mentioned that PJM Load is seemingly made up of aggregate energy consumption in several regions in before the start dates of the individual regions in this analysis (See Table 1). Hence, we can safely assume that the companies in region PJM Load provided electricity to a large number of customers. This aligns with the high average energy consumption throughout the year (See Figure 26). According to PJM, regions in cluster 2 mainly consists of PJM Western sub-regions [15]. With that being said, we believe that Eastern sub-regions of PJM (Cluster 1) have a larger population, hence the higher energy consumption and vice versa for Western sub-regions of PJM (Cluster 2).

It is to be noted that the data on the population served is merely an approximate and is based on the present day. Our time series on the other hand, have different time period. Prior or beyond these time periods, there may be a drastic change in population in each region, merger of companies that may result in greater number of customer served in each region.

To check if we can further groups regions in Cluster 2, we attempted K-means clustering to split the time series into 3 groups.

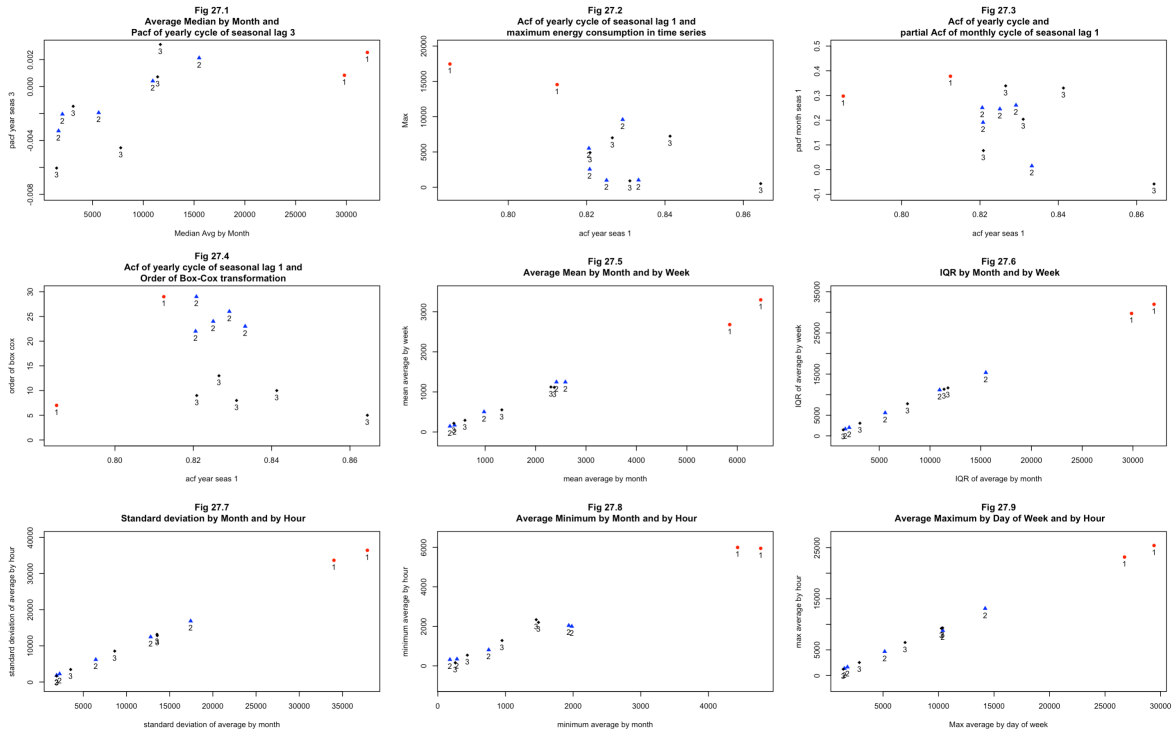


Figure 27: Pairwise scatter plots of features with no distinct means after clustering into 3 groups

We found that when grouping the time series in to 3 groups, there is no distinct groups. When we further analysed the cluster means of the features between the 3 groups, we found that cluster 2 and 3 tend to have relatively similar means. Referring to Figure 27, we do not observe a distinct mean between cluster 2 and 3, which was the purpose of experimenting K-means with 3 groups. The time series in Cluster 2 and 3 appears to be more fitting as a group. With that being said, we decided that performing K-means clustering with 2 groups is more appropriate. It should be noted that the poorly distinguished groups could be a result of the features we chose to use for the clustering process.

# 3 Supervised machine learning

## 3.1 Model Tuning

We started off with a prophet model with default settings. The hour, day of the week, day of the month, week of the year, month, and year were used as features for our initial prophet model.

For our testing, every time series used the same length training and testing data, which are 3 years of training set and 35 weeks of testing data. Because of the varying time period in the data set, some time series have different start and end date of training and testing data set as seen in the table below (See Table 7).

| Time Series | Training Set | Testing Set |
|---|---|---|
| AEP, COMED, DAYTON, DEOK, DOM, DUQ, EKPC, FE, PJME, PJMW | 1/1/2014-12/31/2016 | 1/1/2017-9/2/2017 |
| NI | 1/1/2005-12/31/2007 | 1/1/2008-9/2/2008 |
| PJM_Load | 4/1/1998-12/31/2000 | 1/1/2001-9/2/2001 |

Table 7: Features used in prophet model

For holidays, we tried using common American holidays, such as Christmas, Memorial Day, and Independence Day, but including these holidays did not have a strong effect on our RMSE, or adversely effected our RMSE, so we decided to not include holidays into our calculations.

Additionally, we tried using different trends for our prophet model, which were multiplicative trend and flat trend. Both types of trends increased our RMSE, so we decided to keep the trend as the default additive trend.

Next we began experimenting with the default features. We saw earlier in our EDA that hour, day of the week, week of the year, and month had significant seasonality, while day of the month and year did not. We also noted that week of the year and month were both indicators of annual seasonality. We made models without day of the month and year, and compared models with week of the year versus month. After our testing, we saw that some models would get lower RMSE when those variables were dropped, while some models will get increased RMSE when those variables were dropped. In the end, we went with the simpler model, because the simpler model matches our EDA better, some of our models performed better with the simpler models, and the RMSE loss was not that large for certain models.

| Prophet Model Features | |
|---|---|
| Feature | Description |
| Hour | Hour of the day the observation was taken |
| Month | Month of the observation |
| Day of the Week | Day of the week of the observation |

Table 8: Features used in prophet model

As seen in Table 8, our features were hour, day of the week, and month. Our final model used linear trend and no holiday, and for seasonality, we used daily, weekly, and yearly seasonality, with all of them being set on "auto" except for daily, which was set to TRUE.

## 3.2 Model Results

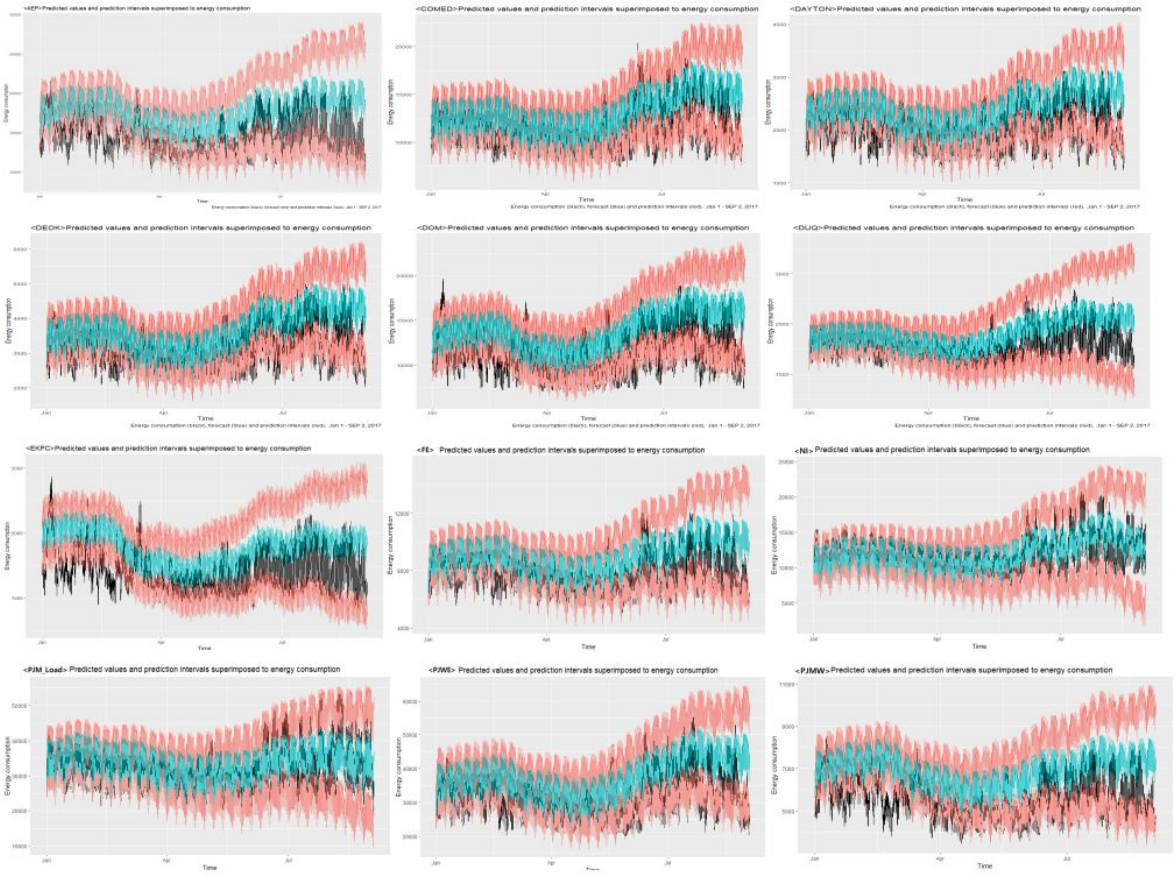

Figure 28: Prediction Plots for all the time series with testing data(Black); Prophet forecast (Green); prediction interval(Red)

Referring to Figure 28, we can see that the further the forecast is in time from the last measurement in the training data, the larger the prediction intervals are, which makes sense in that the closer we are to our training data, the easier it is to predict the range.

| Prophet Model Features | | |
|---|---|---|
| Time Series | RMSE with Final Model | RMSE with Full Model |
| AEP | 3896.659 | 3788.378 |
| COMED | 2727.371 | 2738.456 |
| DAYTON | 527.153 | 527.6951 |
| DEOK | 777.0946 | 753.5888 |
| DOM | 3480.894 | 3680.659 |
| DUQ | **349.0419** | **322.1945** |
| EKPC | 456.5972 | 443.8153 |
| FE | 1779.707 | 1780.305 |
| NI | 1447.827 | 1496.639 |
| PJM Load | 4858.028 | 4974.98 |
| PJME | 8315.651 | 8132.306 |
| PJMW | 1643.903 | 1594.632 |

Table 9: RMSE of all time series using prophet model

Referring to Table 9, we see that the time series with the best RMSE was the DUQ, while the model with the worst RMSE would be PJME. Given that DUQ electricity covers a small region, and that PJME and AEP cover a larger region, we can theorize that our model is better at forecasting electricity companies with smaller region coverage than companies with larger region coverage.

To see how much RMSE was gained or lost with our best time series, we applied the full model to DUQ, and got an RMSE score of 322.195, a difference of approximately 26. This is not a negligible amount of RMSE score lost, which makes us believe that there is information in these basic features not taken advantage of, or features that we could be using that are based off of these features. However, since some of the time series benefit from having the simpler model, it means that we are not completely off the right track either in trying to not use those features.

## 3.3 Best Forecast: DUQ

Figures 29, 30, and 31 display the outputs of the Prophet predictions for the DUQ time series. From Section 3.1, for DUQ, the training period is between 1/1/2014 and 12/31/2016. These time series are predicted 35 weeks into 2018, that is between 1/1/2017 and 9/2/2017.

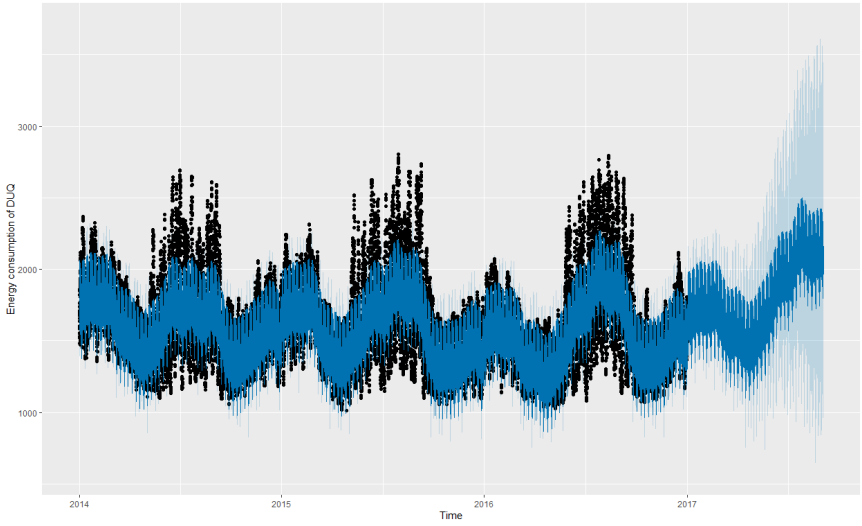


Figure 29: Training Data (Black), Prophet Fit and Forecast (Blue) for DUQ time series

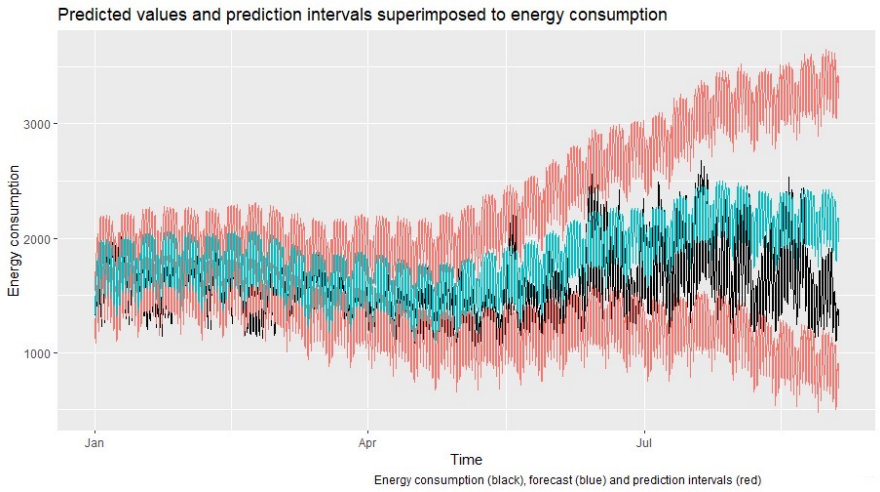


Figure 30: Testing data (Black), Prophet forecasted values (Green) and prediction intervals (Red) for DUQ time series

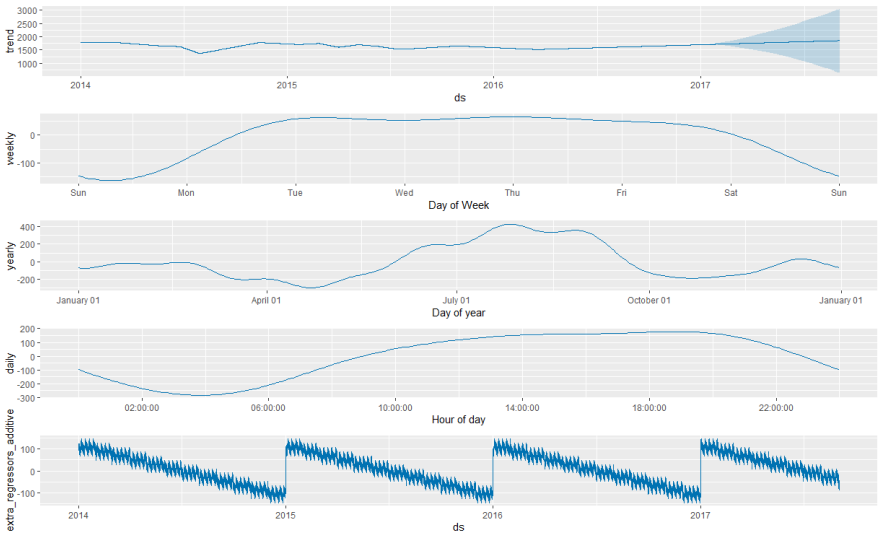


Figure 31: Prophet output breakdown for DUQ time series

Referring to Figure 29, it shows that the prophet fits the training data relatively well. In Figure 30, although our forecasts are not perfect but, we can see that it is relatively close to our testing data. Additionally, it is able to predict the test data of Jan 2017 to March 2017 best and the prediction interval is small. This is expected because when we forecasts further in the future, the accuracy of the forecast decreases. This is reflected in Figure 30 where the prediction interval becomes wider as we forecast further ahead. We can see in Figure 31 that there is a smooth seasonal trend in electricity throughout the day, that there is less energy usage on weekends, and that the amount of energy used has clear seasonality throughout the year.

# 4 Conclusions

Using unsupervised learning with clustering helped us find the biggest distinctions between the electrical regions with little background information. With normal clustering, we would not be able to apply it to time series because we only have one variable, but with feature analysis, k-means clustering is now possible. This means we can find "invisible" trends between different time series data and use them to draw new insights. For an example in the field of animal biology, if we tracked some sort of

bodily rate on a group of animals, such as heartbeats per minute, we can then use clustering with these time series data to see how these animals could be grouped together by their heartbeat patterns.

One advantage of the Prophet model is that without changing the default parameters and without adding features, the model performed quite well. There are a few disadvantages however. Using the prophet model on large time series was GPU-intensive and time-intensive, and as such it was difficult to compare different models, as testing a single change on the model for all of the time series would take more than an hour. Our theoretical backing and EDA was important for refining our model. A shortcoming of our Prophet model was the lack of many non-time features, such as temperature or region size. Another disadvantage was that there was not one best model in terms of RMSE, as some models improved with less features while some models worsened. Prophet's biggest advantage is its robustness and ease of use, which could allow people with little time series knowledge to make robust forecasts.

Overall, these methods have room for a high amount of depth compared to classical time series analysis, because of the introduction of features, the usage of new R packages for coding, and being able to apply previously learned knowledge of statistics to time series.

# References


[1] *American Electric Power*. URL: http://www.aep.com/.

[2] *Applied Energy Services Ohio*. URL: https://www.aes-ohio.com.

[3] *Atlantic City Electric*. URL: https://www.antlanticcityelectric.com/.

[4] *Baltimore Gas and Electric Company*. URL: https://www.bge.com/.

[5] *Commonwealth Edison Company*. URL: http://www.comed.com/.

[6] *Delmarva Power*. URL: https://www.delmarva.com/.

[7] *Dominion Energy*. URL: https://www.dominionenergy.com/.

[8] *Duke Energy*. URL: https://www.duke-energy.com/.

[9] *Duquesne Light Company*. URL: https://www.duquesnelight.com/.

[10] *East Kentucky Power Cooperative*. URL: https://www.ekpc.com/.

[11] *First Energy Corp*. URL: https://www.firstenergycorp.com/.

[12] Rob Mulla. *Hourly Energy Consumption*. Aug. 2018. URL: https://www.kaggle.com/robikscube/hourly-energy-consumption.

[13] *Orange and Rockland Utilities*. URL: https://www.oru.com/.

[14] *PECO Energy Company*. URL: https://www.peco.com/.

[15] *Pennsylvania-New Jersey-Maryland Interconnection*. URL: https://www.pjm.com/.

[16] *Potomac Electric Power Company*. URL: https://www.pepco.com/.

[17] *PPL Electric Utilities*. URL: https://www.pplelectric.com/.

[18] *Public Service Enterprise Group*. URL: https://corporate.pseg.com/.

[19] *U.S. energy information administration*. URL: http://www.eia.gov/.


# Acknowledgments


We consulted Professor Sanchez's Stat 170 class lectures throughout this assignment. To analyze the dygraphs and seasonal box plots of varying cycles, we used concepts covered in Week 9 lectures "Kaggle: hourly energy consumption. Time Series with multiple frequencies". To generate the plots, we referred to the videos and accompanying R script from the same Week 9 lectures and "3/1/2022 Lecture Video of the discussion"

To select the features for the K-means clustering based on features, we referred to inputs that the class put in a google document "Time Series features". We also referred to Week 8 lectures "Dr. Sanchez's in person/zoom lecture of 2/24/2022 -second segment, which contains only discussion of the google doc due 2/24 on features.".

To perform and analyse the results from the K-means clustering, we applied concepts covered in Week 8 lectures "Time series features engineering 1." and referred to the R script accompanying the lecture video "Program used in the first segment of today's lecture to do the clustering of the ACF and PACF features of several time series".

Last but not least, to fit a Prophet models to the time series and conduct the analysis, we applied concepts covered in Week 8 lectures "When additional customized features must be taken into account" and referred to the R script accompanying the lecture video " Dr. Sanchez's in person/zoom lecture of 2/24/2022- Segment 3, the main lecture". We also referred to a comment about prophet by Dr. Sanchez in the discussion board 'HWK4 Q&A -Discussion among students about homework 4-groups help groups'. We also used the following video to help with our prophet modeling: https://www.youtube.com/watch?v=3hh7XO9aFBAa