Time Series Analysis of Hourly Energy Consumption

The Problem

Proper allocation of energy is a vital part of avoiding minor power inconveniences and nation-wide blackouts. Electrical utility companies must diligently plan ahead and properly allocate energy across their generating units to meet their regional energy demand. If there is a significant imbalance between energy demand and energy output, this could lead to blackouts consequently economic loss. In addition, it is important to not allocate too much energy as this could lead to electric waste. The ability to forecast the energy consumption will allow these companies--present across the nation-- to always be prepared for any possible energy imbalances.

The Data

PJM Interconnection LLC (PJM) is a regional transmission organization (RTO) in the United States. It is part of the Eastern Interconnection grid operating an electric transmission system serving all or parts of Delaware, Illinois, Indiana, Kentucky, Maryland, Michigan, New Jersey, North Carolina, Ohio, Pennsylvania, Tennessee, Virginia, West Virginia, and the District of Columbia.

The Data contains hourly energy consumption --in Megawatts--data from 2002 to 2018 covering the following regions: Delaware, Maryland, New Jersey, North Carolina, Pennsylvania, Virginia and District of Columbia.

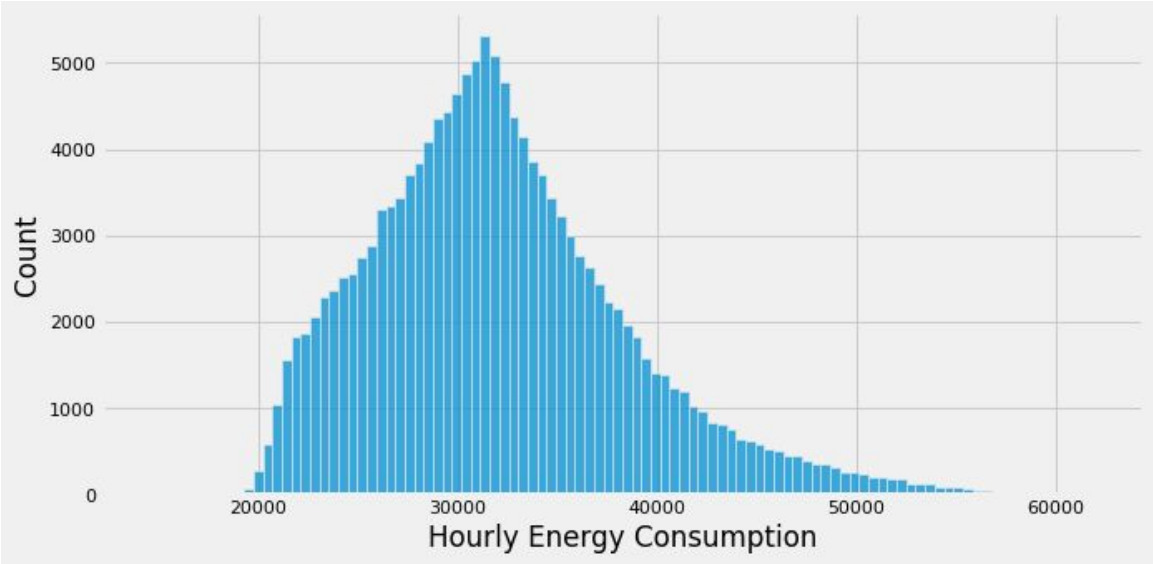
Data Preparation

The data was processed to be usable in a time series analysis. This involves proper reading of the index such that it's a datetime object. Other general cleaning processes involved, sorting, removing duplicates, renaming columns, and forward filling any null values. As I moved forward through the analysis I also added more features that describe yearly, monthly, weekly (and etc.) trends. This was for both visualization purposes as well as modeling purposes mainly pertaining to seasonality.

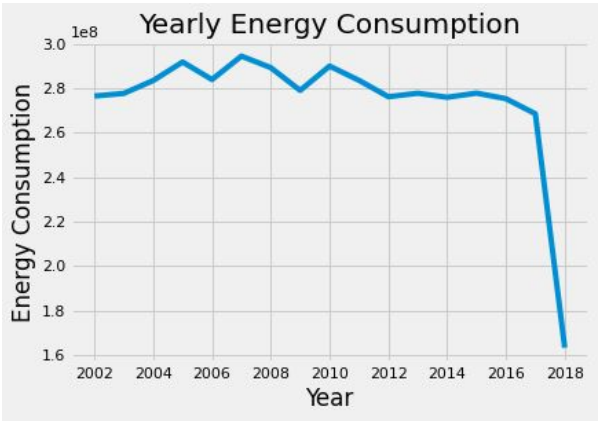
Exploratory Data Analysis

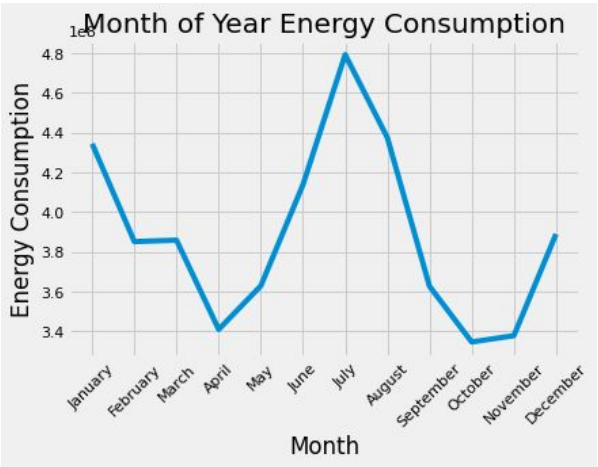
In this section I explored the data with the goal of finding interesting patterns. Some of these patterns were quite clear on a yearly, monthly, and weekly basis. Let's first take a look at the distribution of Energy consumption (EC).

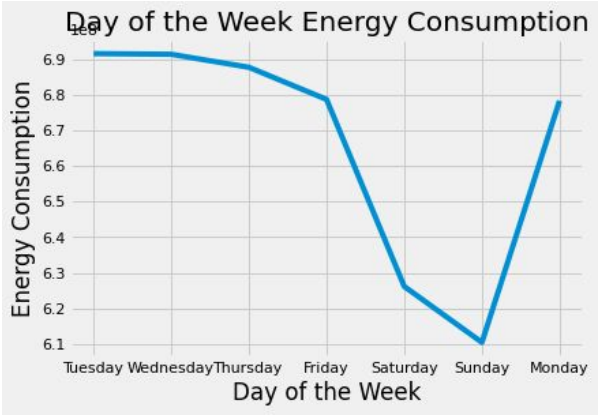
Below, we can see that it has a slightly right skewed distribution (somewhat normal) with a mean of 32,079 MWs (with a 6,464 standard deviation). This value will be useful when we later evaluate our models because it is rather large. Skew is explained by the fact that people try to use energy as efficiently as possible, at least in most cases.

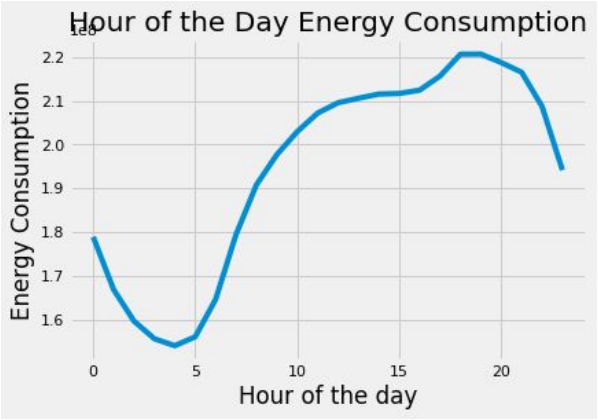


Below we have some insights about seasonal patterns.



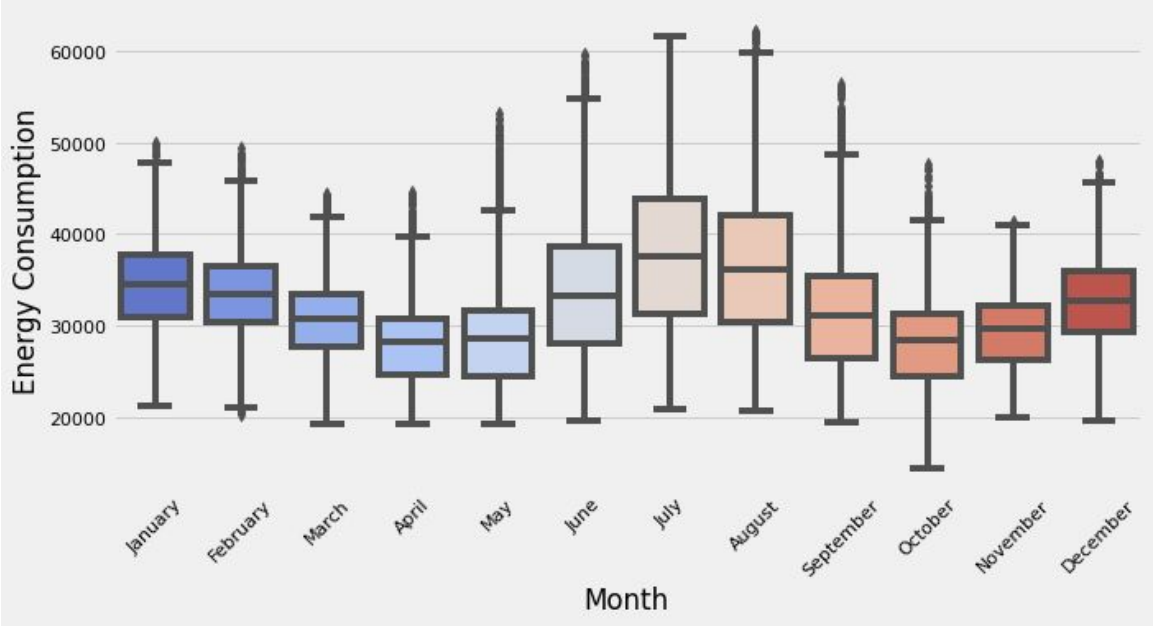


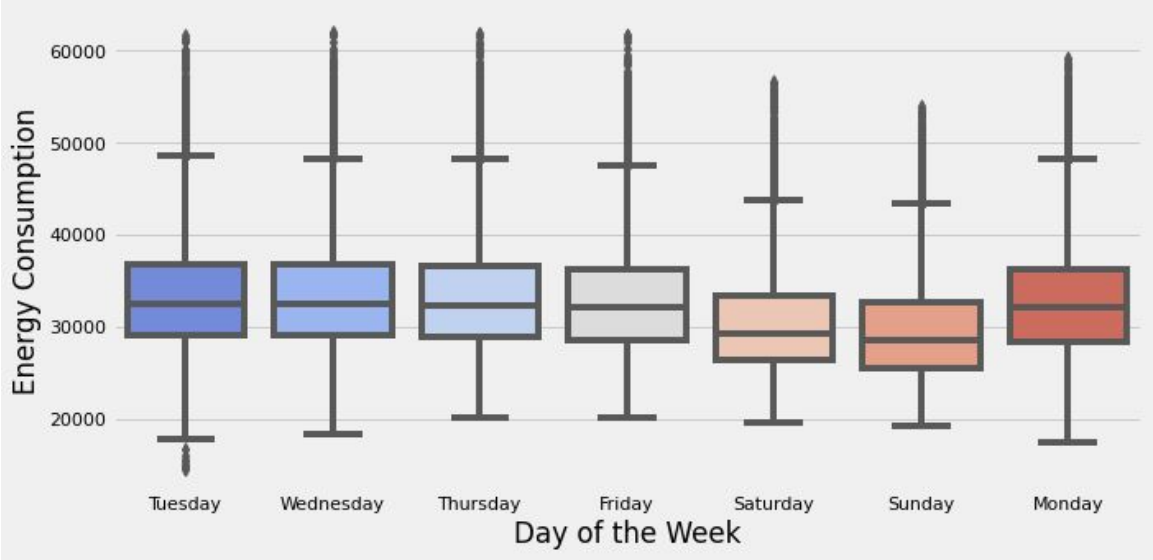




These are the most notable seasonal features.

1. The yearly EC is constant up to 2015, a decreasing trend from 2015 to 2017.
2. Monthly EC is high during the summer --peaking in July due to air conditioning? -- summers months should be closely monitored by the energy companies. Significantly high energy consumption between January and December perhaps due to holidays? .
3. Day of the week, we can see that the energy consumption drops during weekends, perhaps because certain companies do not function during weekends that use up significant energy?
4. The hourly EC shows increase starting 6 A.M. to 8 P.M. (20 in military time) then predictably is at lowest during the "sleep" times.
5. Week of year, day of year, and quarterly EC show the same pattern as the monthly distribution (i.e. highest during summer).

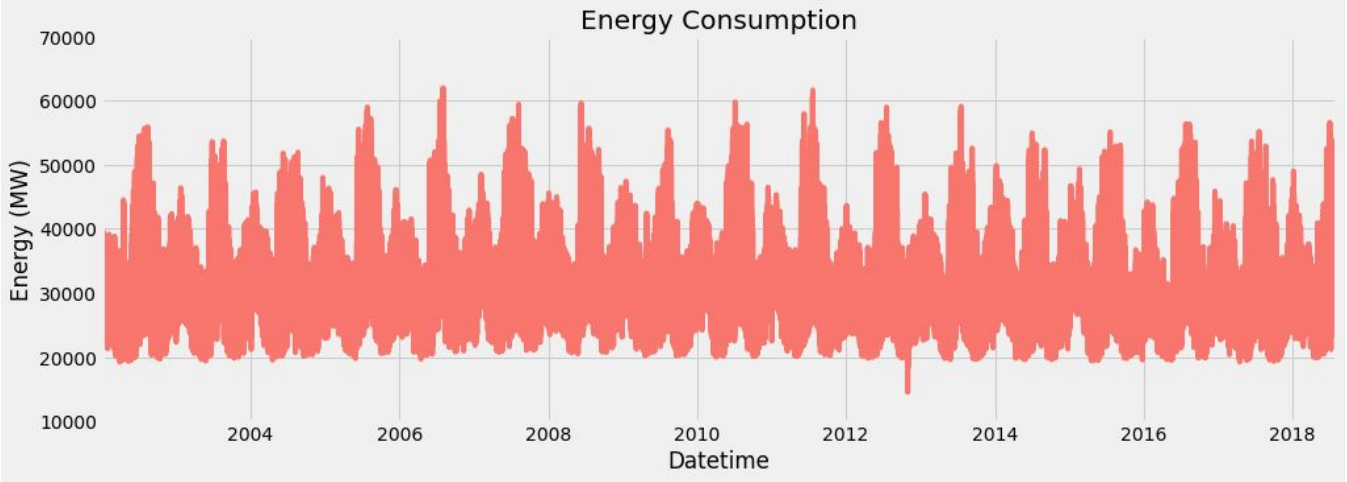




We can see that the distribution(the dispersion) of energy consumption is quite the same on a daily basis; however, these differences are much more significant on a monthly basis. Thus, we have a larger spread of energy consumption in June, July, and August possibly due to the summer's heat.

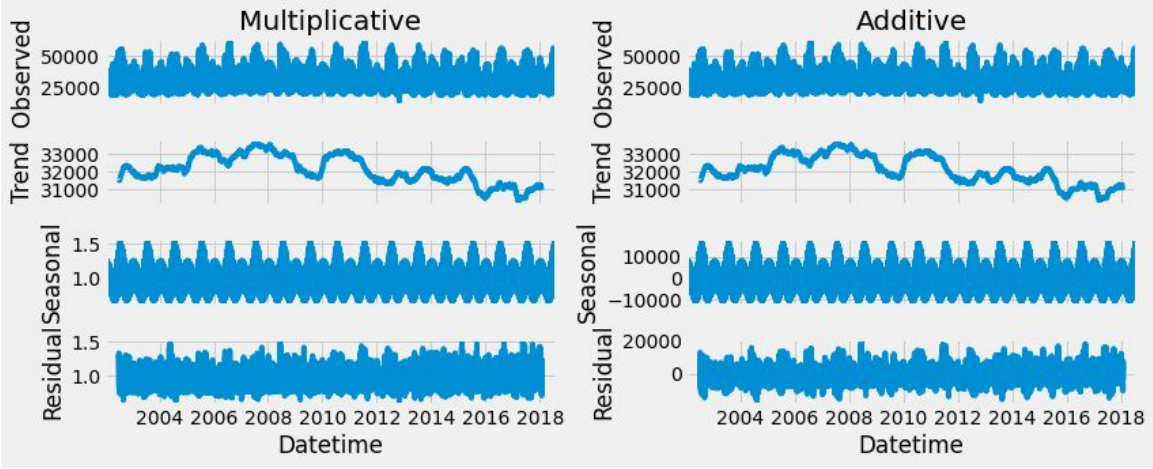
Data Analysis

Let's take a look at our time series data and its components and ensure that it's stationary.

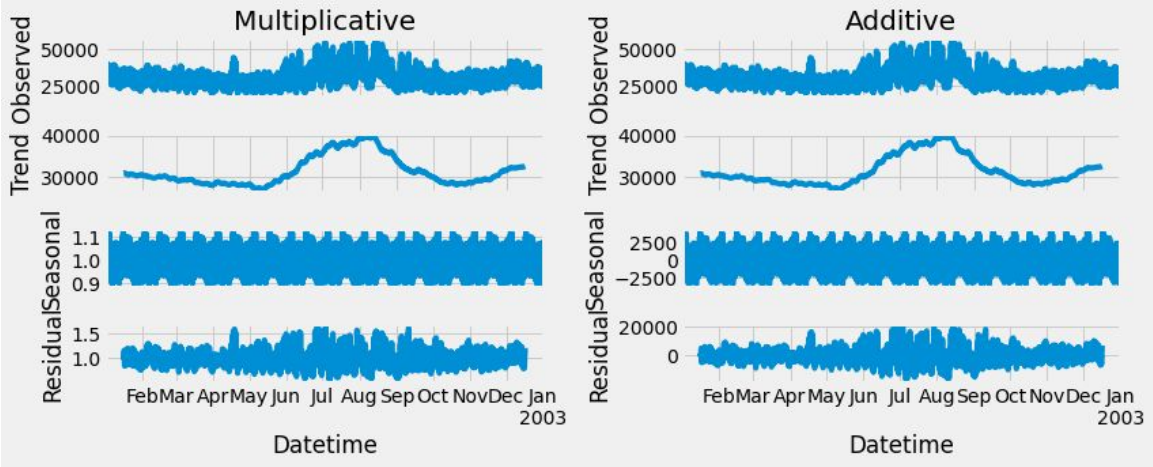


On the surface, the series looks stationary. We can also note that the energy consumption at the end of October 2012 is relatively low. This is because Hurricane Sandy caused significant destruction (70 billion USD). Let's now take a closer look at the components of our series : Trend, Seasonality, Residual.

Let's explore our components based on yearly seasonality first.



We observe univariant cyclical patterns for both models thus our data is yearly seasonal. Let's do the same for monthly seasonality.



We observe a univariate pattern here as well so our data is also monthly seasonal.

Stationarity

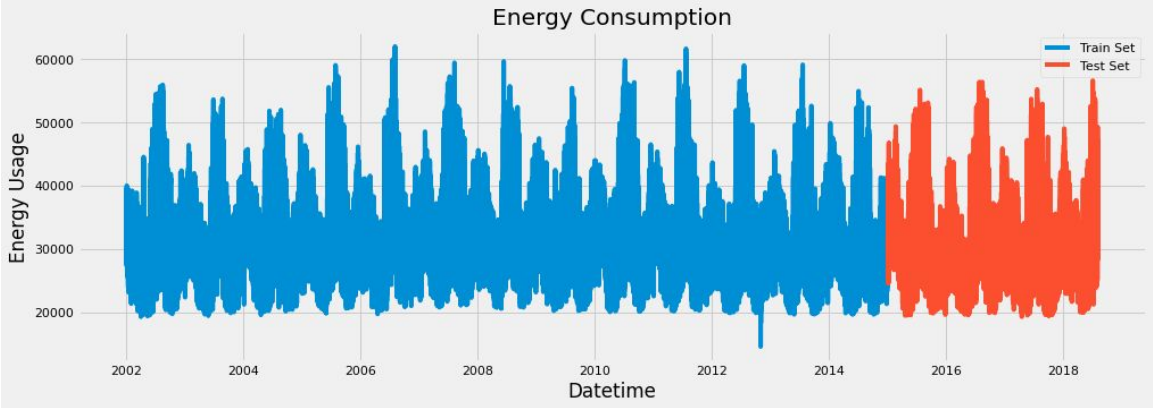
Definition : Intuitively, stationarity means that the statistical properties of the process do not change over time. This is an important assumption in a time series analysis. I will use the Augmented **Dickey Fuller** test to ensure the data meets the stationarity assumption.

Null Hypothesis	Alternate Hypothesis	Critical Value (based on alpha = 0.01)	Test Value
Unit root is present in our series	Our series is stationary	-3.43	-19.97

Since our Test value < Critical, we fall in the rejection region. Thus we reject the null hypothesis. This means that there is statistically significant evidence to show that our series is in fact stationary.

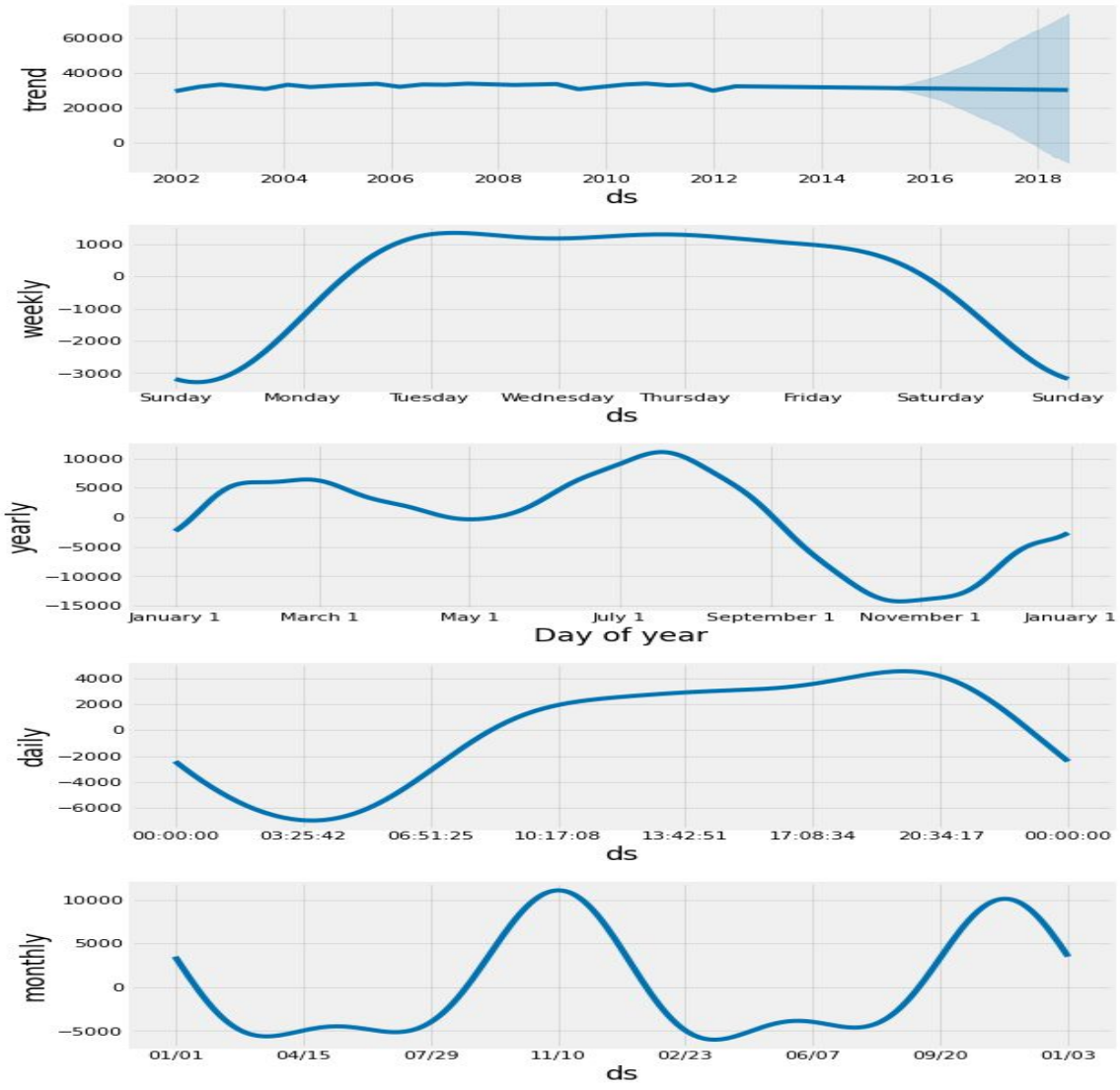
Times Series Forecasting

I split my data into training and test sets using the last 3 years as the latter. (about 80/20)

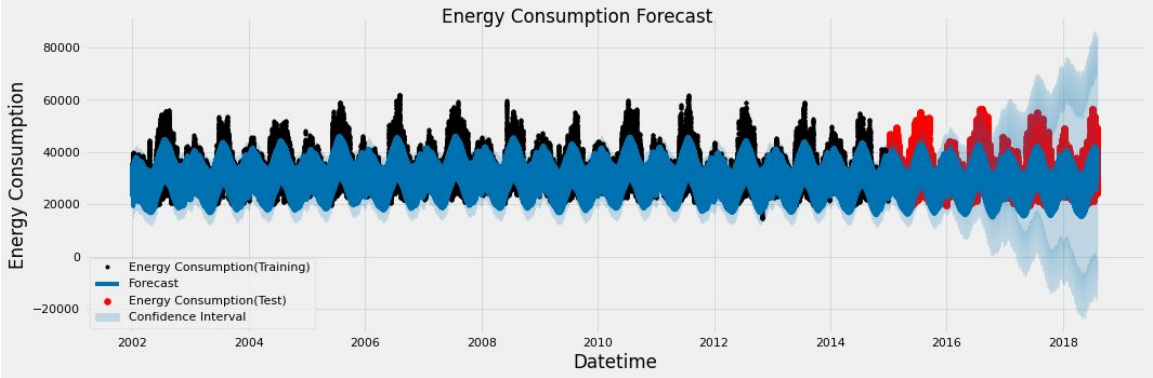


I. Model 1: FB Prophet

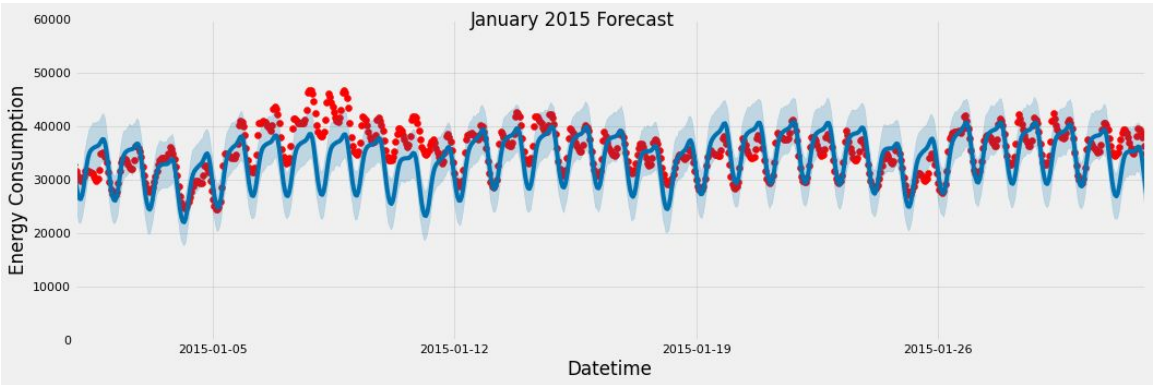
Prophet is a framework created by facebook to forecast time series by decomposing the series into its components. As we can see below, the framework has been able to recognize all the seasonalities, and separated the trend.

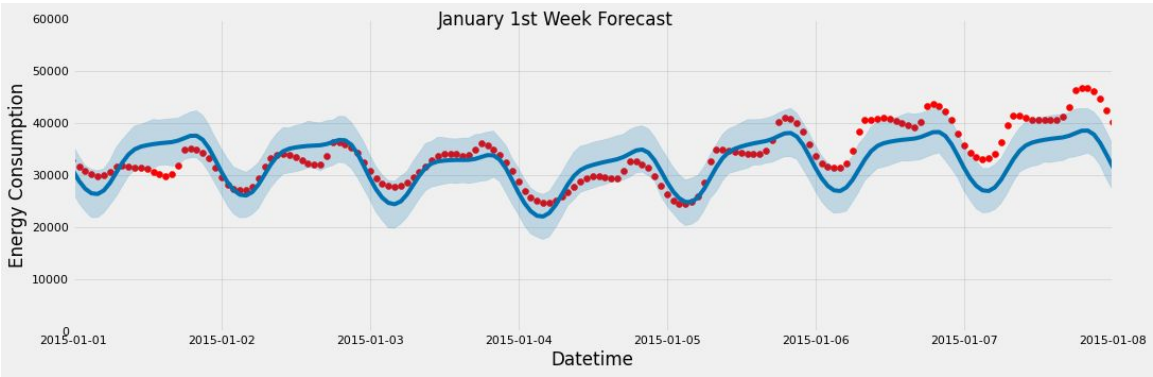


Here are the forecasting results with the prophet model which also shows confidence interval of the given forecast.

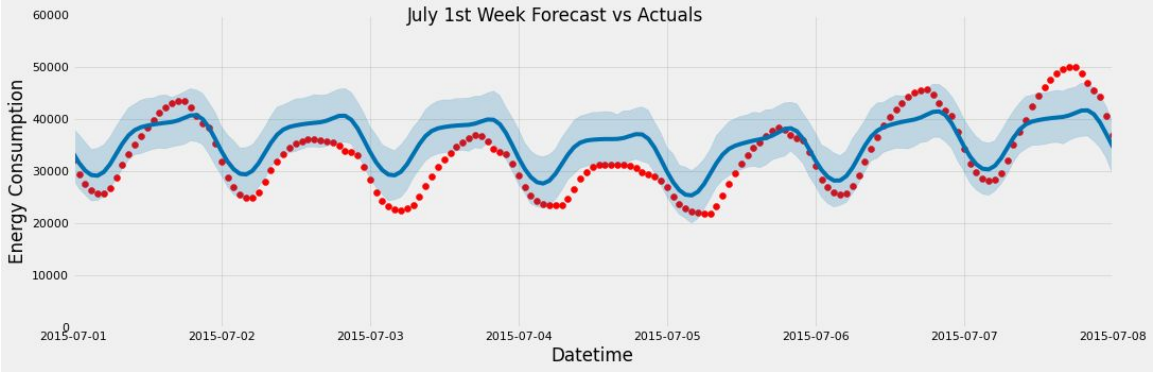


The model does a good job at following the general direction of the test data, but fails to reach the peaks consistently. This is expected because those peaks are quite random and it'd be difficult to fully model them without overfitting.





There is an example of such a peak here in January. Within 95% confidence interval; however, the model almost captures the real values most of the time. Let's take a look at July.



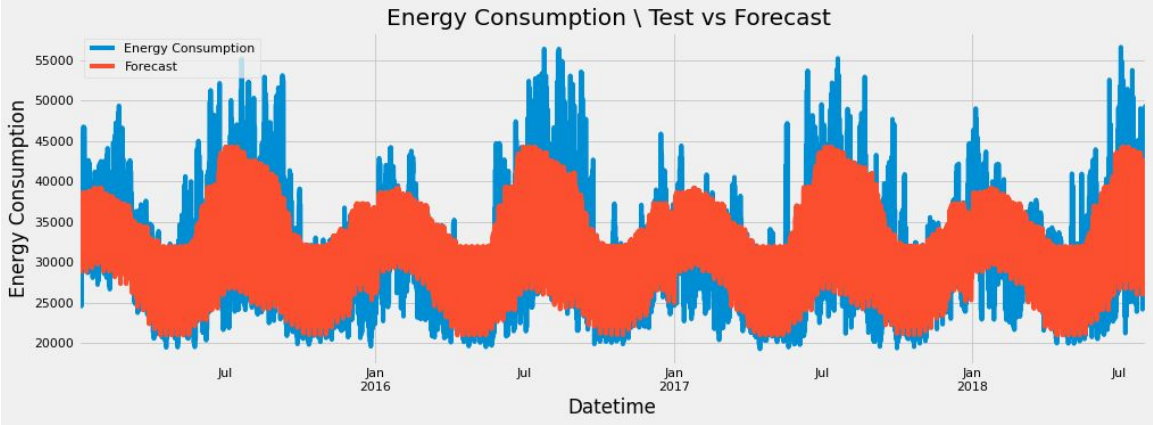
The model overestimates in some instances here, while underestimating in other instances. Generally speaking, it would do a good enough job in terms of assisting decision making, but it can't definitely be improved.

We'll compare error metrics after discussing our next model.

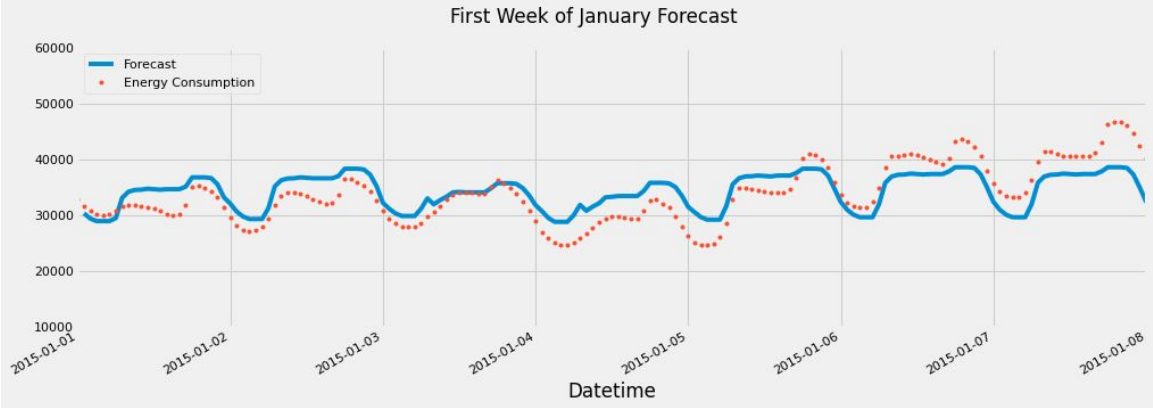
I. Model 2: XGBoost

XGBoost is an implementation of a gradient boosting framework, where new trees fix errors of those trees before them until no improvements can be made. We'll look at some visualization regarding the forecast. We will discuss the metrics for both models in the end of our report.

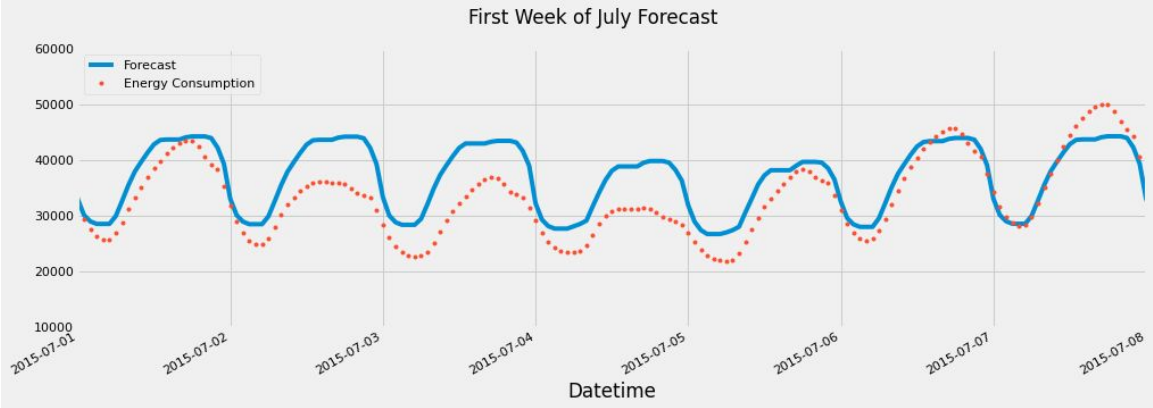
Here are the forecasting results with the XGBoost model on our test data.



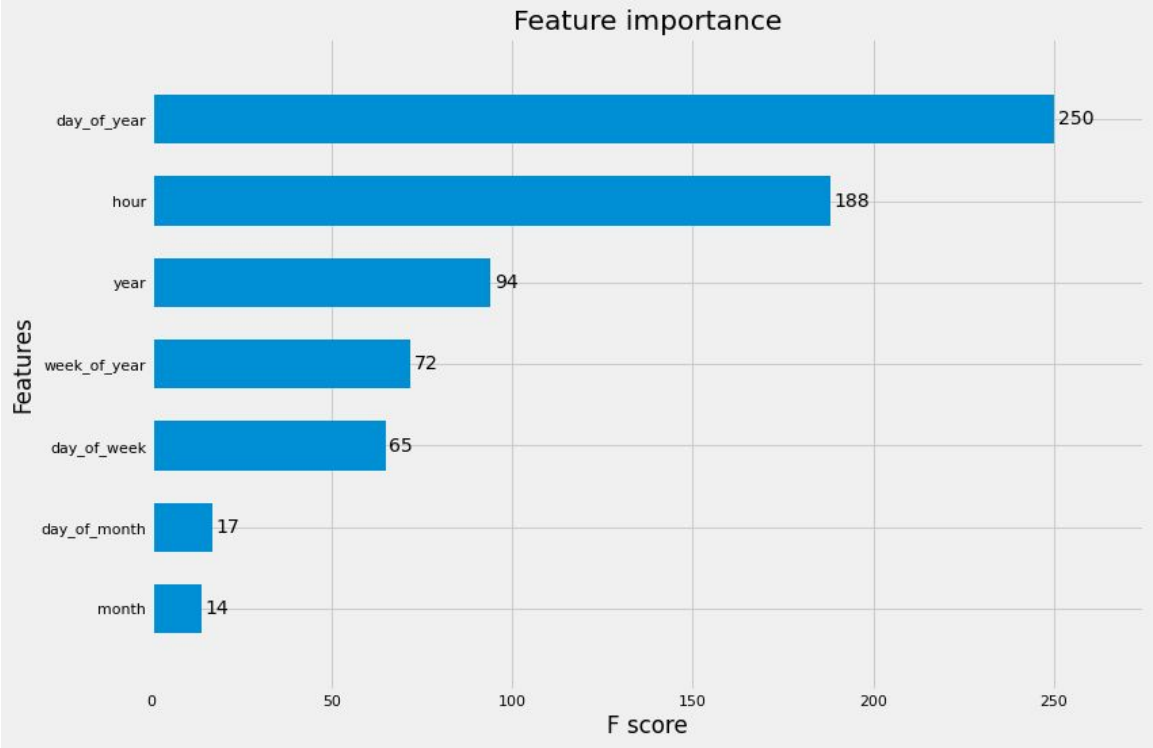
We can see that the XGBoost model does a decent job on our test data. It's hard to tell how well the model is doing without metrics. Let's take a look at how our model does for specific weeks/months before looking at metrics.



For the first week of January, we can see that it follows the general pattern quite well. It's not fully accurate, but that's important as we would not want our model to overfit.



Looking at one of the most important months, our model does a good job not underestimating the energy consumption. This is important because we would not want to have shortage in allocation of energy in the middle of the summer.



From the visualization above we can see that day of the year plays the most important role in terms of forecasting energy consumption followed by the hourly energy consumption.

Error Metrics

We'll use the following evaluation metrics:

1. Mean Squared Error
2. Root Mean Squared Error
3. Mean Absolute Error
4. Mean Absolute Percentage Error

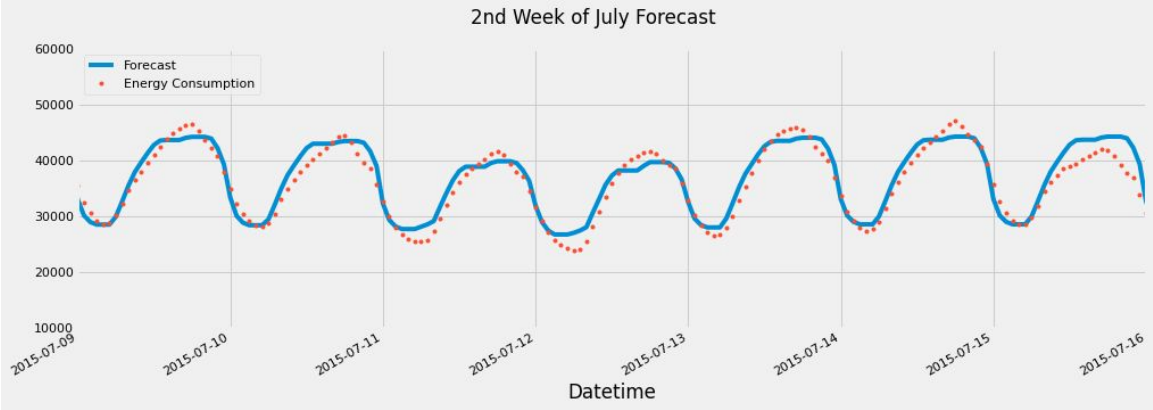
We'll focus on 2, because they're the most intuitive.

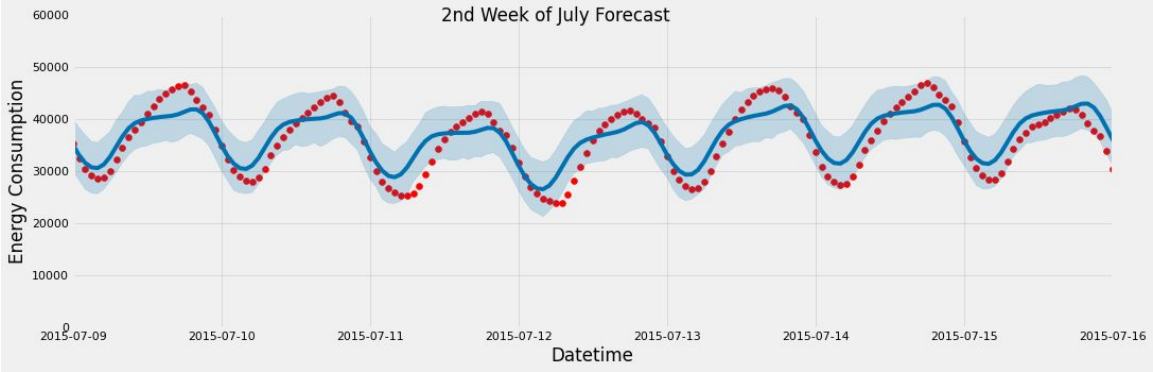
MAPE helps us understand % error on the absolute value and is indifferent to the size of the time series while RMSE generates a value that is relative to the Standard Deviation of our original series.

Model	**RMSE**	**MAPE**

Prophet	4195.90	9.86%
XGBoost	3745.90	9.11%

The results are unsurprisingly similar with a slight edge to the XGBoost model. The RMSE of around 4000 KWs is not significant because the standard deviation of the energy consumption was approximately 6464 KWs. For some visual comparison, we can see that both of these models do well predicting a random week of July.





Conclusion and Future Improvements

The final goal for this project was to develop a forecasting model that can be utilized by the electrical utilities to effectively and efficiently plan their energy generation operations. Such forecasts can be very useful for the utilities in planning their day to day operations, meeting energy demand, and avoiding any excess generation of energy.

We were able to extract general patterns in our EDA. We've seen that energy consumption is significantly higher during summer due to higher temperatures. We've also observed similar patterns for weekdays and day times. In addition, we were able to identify multiple seasonalities in our data and create 2 accurate models to forecast energy consumption in the East region. Such models will be useful tools for electrical companies in managing their resources whether it means saving them from blackouts or electrical waste.

In order to improve my model, it would be useful to add holiday indicators and weather data sources. The addition of such information could possibly make our models slightly more accurate. Perhaps, we would be able to forecast those extreme consumptions more accurately than not. I am also curious to see how well a simpler algorithm such as ARIMA does with this data.