Forecasting Energy Usage in Appliances and Lights

Jessica Parks

April 2023

Introduction

Topic Focus

Now more than ever companies are having conversations about creating more sustainable products. We see this in many areas of business. Clothing is being advertised as made with sustainable materials[1]. The automotive industry is increasing investments into the research and development of electric cars to reduce fossil fuel emissions[2]. Better infrastructure is being built to achieve LEED certification[3]. Not only is this good for our environment and communities, but from a marketing stand point, it is good advertisement for businesses too.

It is one thing for a company to highlight its recent sustainability achievements but it is another to show how they have maintained this effort and improved upon it overtime. As understanding environmental impact has grown in awareness, some companies have embellished their efforts for advertisement rather than transparency. This has become such an important issue that a term has even been created for this, "greenwashing". This is not the reputation a business wants to have. In this case, it is important that companies make a collective effort for their business model and their consumers to reduce their environmental impact.

How can a business prove to their competitors and their consumers they are working toward reducing their carbon footprint? By showing the improvements they have made in their past, along with the trajectory they are on for their future. Even if a business' current trajectory is not the direction they want to be on, they can show interventions they are currently making to mitigate their impact. This is something everyone can feel good about. Their employees will feel good about their efforts being made for the environment and future generations, along with the consumers who know they are purchasing a product that was consciously made to minimize their energy usage and, ultimately, the cost of using it.

Purpose of Study

Producers of appliances and lights understand that not only is energy efficiency good for the environment, but they are also important to the consumer. With more understanding around energy output of objects found in a home, customers are less likely to spend money on a product that requires a lot of energy, especially if the energy is wasted and dispelled to the surrounding environment.

The energy output of appliances and lights is not an isolated output. This can be dependent upon many features such as the materials the product is made of, the size of the product, etc. Therefore, when predicting the energy output, it is important to take other features into consideration when possible.

In our case, we will also look at room temperature, room relative humidity, and typical environmental factors (dewpoint, outdoor temperature, etc.). While these are not the only factors, they are the ones that were provided in our dataset. The rooms in the house do not all have the same number of appliances nor are all the appliances in different rooms outputting the same amount of energy.

Task Definition

The primary objective of this paper is to explore and analyze the energy consumption patterns of various appliances and lights within residential households over a period of four months. To achieve this objective, we will examine a multitude of factors that could potentially impact the energy consumption of these devices, including the temperature and relative humidity of individual rooms, as well as the prevailing weather conditions in the environment at the time. Our model will identify the significant features that impact the energy output of both appliances and lights. By analyzing these various factors, we aim to develop a comprehensive understanding of the complex relationship between environmental factors and energy consumption. In addition, we will leverage this understanding to build a multivariate timeseries model that can accurately forecast the energy consumption of appliances and lights in the future. Ultimately, the results of this study could provide valuable insights into the factors that impact energy consumption patterns, and inform the development of more efficient and sustainable energy consumption strategies in the residential sector.

Methodology

Data Collection

For background information on the data, this was collected in 2017 in Belgium and contains nearly 20,000 data instances. Data was collected every 10 minutes over the course of 4.5 months. Each room's humidity and temperature were recorded, along with the energy output of the appliances and the lights. The devices used in this data collection included wireless sensors for each room and an energy meter for the appliances. Outdoor data was also collected from a nearby airport weather station to where the home was located; the outdoor data includes the temperature outside (Celsius), the air pressure (mmHG), the relative humidity (%), wind speed (m/s), visibility (km), and dewpoint (Â°C).

This dataset is important for addressing the energy usage issue because it does not treat energy output as an isolated event. By analyzing not just indoor temperature and relative humidity and outdoor factors, we will be able to have a better idea of the energy usage of appliances and lights.

Algorithm Definition

There are a few features of forecasting that make timeseries analysis different from other predictive modeling techniques. These are essential to maintain when beginning data analysis. One distinguishing, but significant feature about timeseries analysis than other, regression models is the relationship between the datapoints. Each point is dependent upon its past, ordered

points. Therefore, this structure of historical data must be maintained. We cannot shuffle the data and test it. Similarly, the datapoints are not independent of one another like they are in other regression models. There is a certain structure to timeseries data that must be maintained when carrying out our methods.

For this reason, there are many models that are specific to timeseries analysis. These depend on important features, that will be discussed later on. Since we are developing a multivariate model, this did restrict the timeseries specific models we could use down a bit. There are multivariate models that have been constructed with a univariate model foundation. We will be considering a model of this kind called the Vector Autoregressive Model (VAR). This multivariate model is an expansion of the univariate model called the Autoregressive Model (AR). This model was chosen due to its ability to be flexible, easy to use and successful algorithms for multivariate forecasting[4]. In this type of model, the dependent variable is a linear function of past values of the dependent variable. This model can also use a certain number of lags used to predict its current value. It is a relatively straightforward approach to a model:

$$y_1(t) = a_1 + w_1 11 \cdot y_1(t-1) + w_1 12 \cdot y_2(t-1) + e_1(t-1)$$

Multivariate timeseries forecasting models, such as the Vector Autoregressive Model (VAR), can be powerful tools for predicting future outcomes based on multiple related variables. However, in order to develop an effective model, it is important to carefully examine the underlying data and ensure that it meets certain criteria. This includes evaluating the stationarity of points, identifying seasonality and trends, and checking for the best lag value. This will be discussed more in *Prepare the Data and Tuning.*

Exploratory Data Methods

Before looking at patterns and trends that are specific to timeseries analysis, we need to first look at some basic data and relationships. For example, in the graphs of Figure 1 below, we can see the trends of energy usage by appliances and lights in the right hand column. In the left column, we can see the temperature and relative humidity of the kitchen (chosen since many appliances are in this one room). Based on the graphs, we can see there is an overall decline in trend as time goes on. Look more closely though, and we see these spikes and declines are not all the same. The graphs of the appliances and lights have a greater amount of spikes and variations. This indicates it is important to look more closely at the relationship between these variables later on.

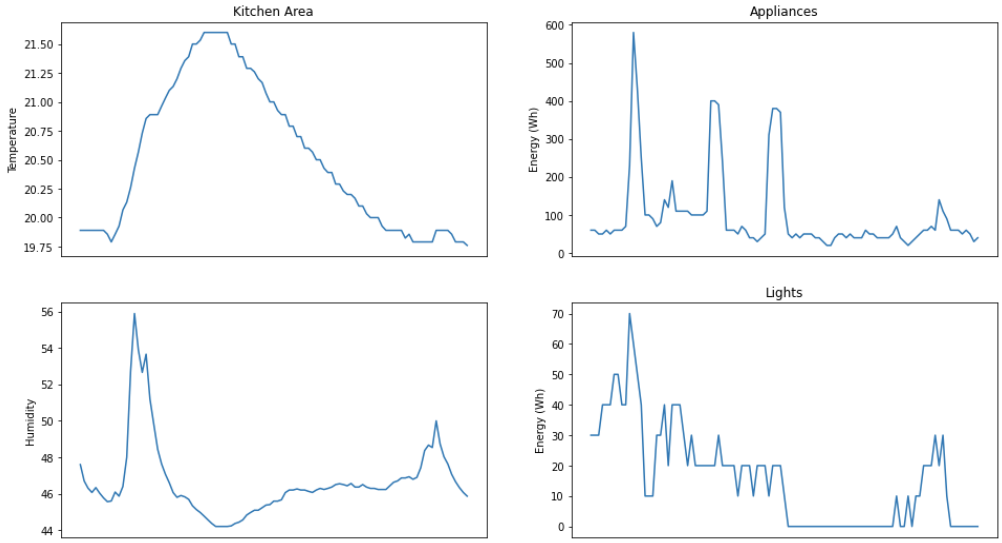

Figure 1

While this is helpful to see that there is potentially a relationship between these variables, this is not enough information to begin applying our data to the VAR algorithm yet. Our model aims to quantify how much the appliances and lights are correlated with the relative humidity and temperature. In order for us to forecast their energy consumptions into the future, we would like to know how much each of these variables will impact our model. This is shown in Figure 2 below in our correlation plots. On the left, we see the correlation between relative humidity, appliances and lights; on the right is the correlation between temperature, appliances, and lights. Each number next to T and RH_ represents the assigned number to each room of the house (I.e. T1 and RH_1 are the temperature and the relative humidity of the kitchen.

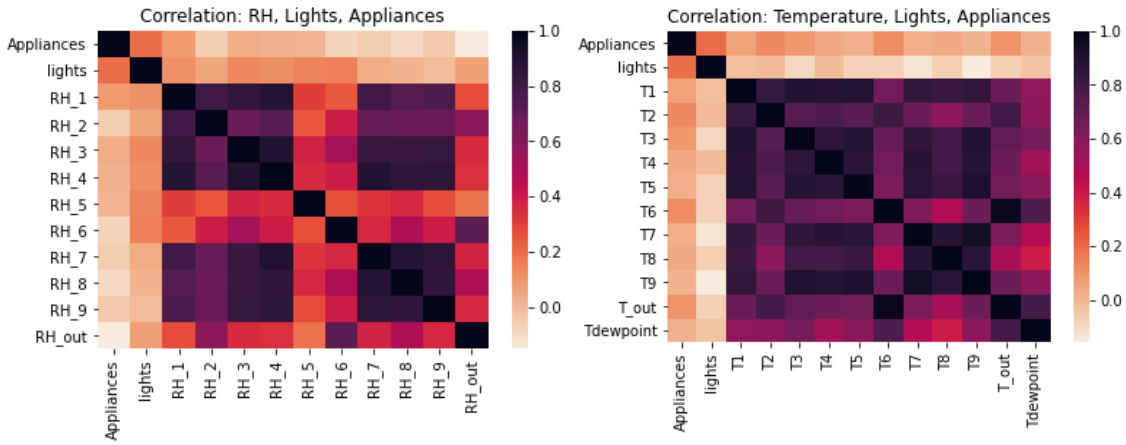

Figure 2

Upon first glance, these correlation plots reveal a few interesting things. It appears there is not as much of a correlation between temperature and the appliance and lights as there is for relative humidity and appliances and lights. Just from these two plots, we hypothesize that relative

humidity has a stronger impact on the energy usage of lights rather than appliances. On the flipside, we hypothesize temperature has a stronger impact on the energy usage of appliances than lights.

As mentioned earlier, there are some important features of timeseries analysis that make it different from other regressions and therefore, we need to look at other factors that could affect our VAR model. One key consideration is the stationarity of the data. Stationarity refers to the statistical properties of the timeseries remaining constant over time, such as the mean and standard deviation. If the data is non-stationary, it may exhibit trends or other patterns that can make forecasting more challenging. One quick way to look for stationarity is to graph the rolling mean and standard deviation of the response variable and see if it fluctuates over time. If so, then this can indicate non-stationary data. Figure 3 below shows the graphs made for the appliance and lights. As shown below, there is clearly fluctuation in the rolling mean and standard deviation of both variables, indicating they are non-stationary or the features that impact them are non-stationary. This will be looked at more closely in the next section.

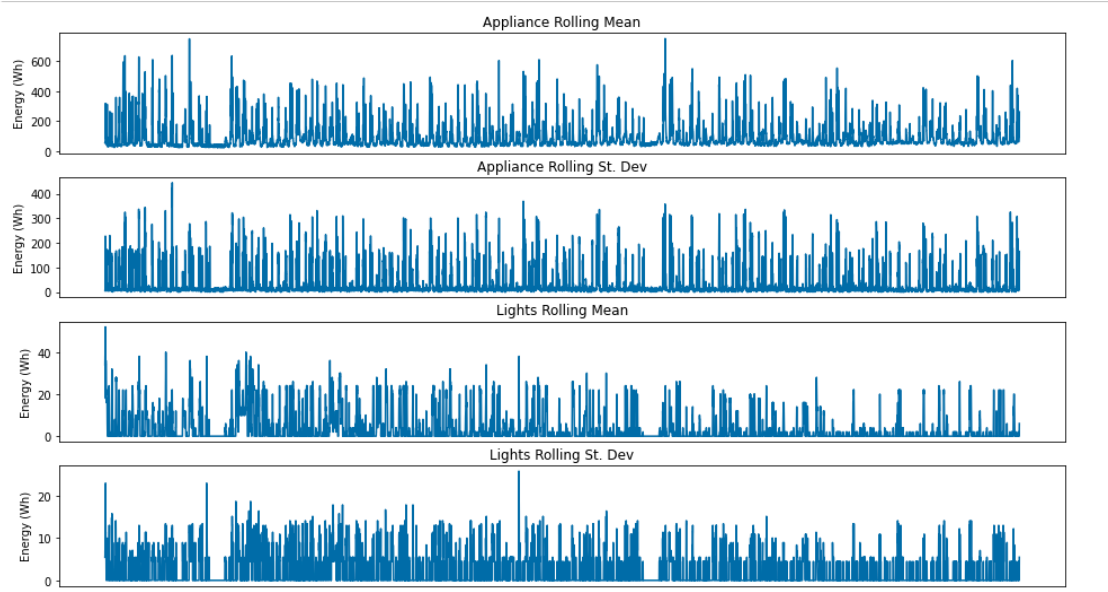

Figure 3

Prepare the Data and Tuning

There is quite a bit that should be checked in timeseries data before applying the selected model to it. This step is important to ensure that the data we pass through the model is stripped of as much extraneous noise as possible and so the model produced is representative of the underlying patterns in our data.

While we briefly looked at correlation in our exploratory data analysis, it is important to take a closer look at how the correlation is occurring within variables. Since we were working with Appliance and Lights, I constructed ACF and PACF plots for both of these variables (Figure 4)

to gain a better understanding of their correlation with themself and their lag values. We can see there is some correlation with a data point and certain lag values.

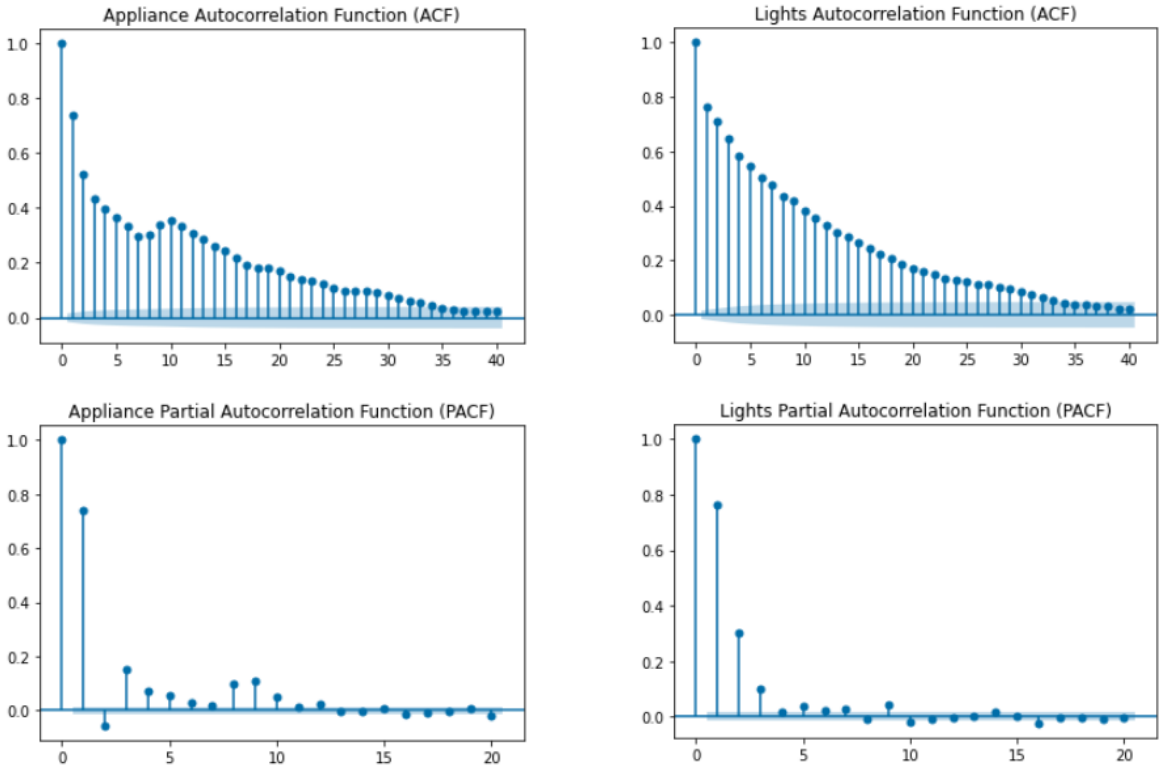

Figure 4

Since we saw there was a fluctuation with our data's rolling means and standard deviations, we wanted to have a quantitative understanding of the fluctuation. There are statistical tests that can be used to identify if a dataset shows stationarity or non-stationarity. In order to test for this, we applied the Augmented Dickey-Fuller (ADF) test[5] to each of our variables.

$$\Delta y_t = \alpha + \delta t + \beta y_{t-1} + \sum_{j=1}^{p} \rho j \Delta y_{t-j} + \varepsilon_t, \quad t = 1, \ldots, T$$

This test examines whether the first-differenced time series is stationary, which means it subtracts the value of each observation from the preceding observation and therefore factors in lag values. If the resulting time series is stationary, it implies that the original time series was non-stationary. When calculating this, we deemed a variable statistically significant if its p-value was less than or equal to 0.05. When we ran the ADF test on each of our variables, there were seven variables that were deemed not stationary. As a result, we needed to difference our data; this helps stabilize the changes that occur in the mean and standard deviation, thus reducing the trend and seasonality of the data.

Seasonality is another important factor to consider in timeseries analysis. Many timeseries exhibit recurring patterns over a fixed period of time, such as daily, weekly, or monthly cycles. Since our data is working with weather related features, one would hypothesize there is

seasonality in our data. Identifying and accounting for these patterns can improve the accuracy of the forecast. In order to check for our seasonality, we used seasonal decomposition, which separates the data into three components: trend, seasonality, and residuals. When considering the seasonality of appliances and lights below (respectively Figure 5, Figure 6), we can see the seasonal trend we expected that occurs each day. For example, in the second section we can see the decomposition of lights; one would expect lights to not be on as much during the day and on more at night when it is dark out.

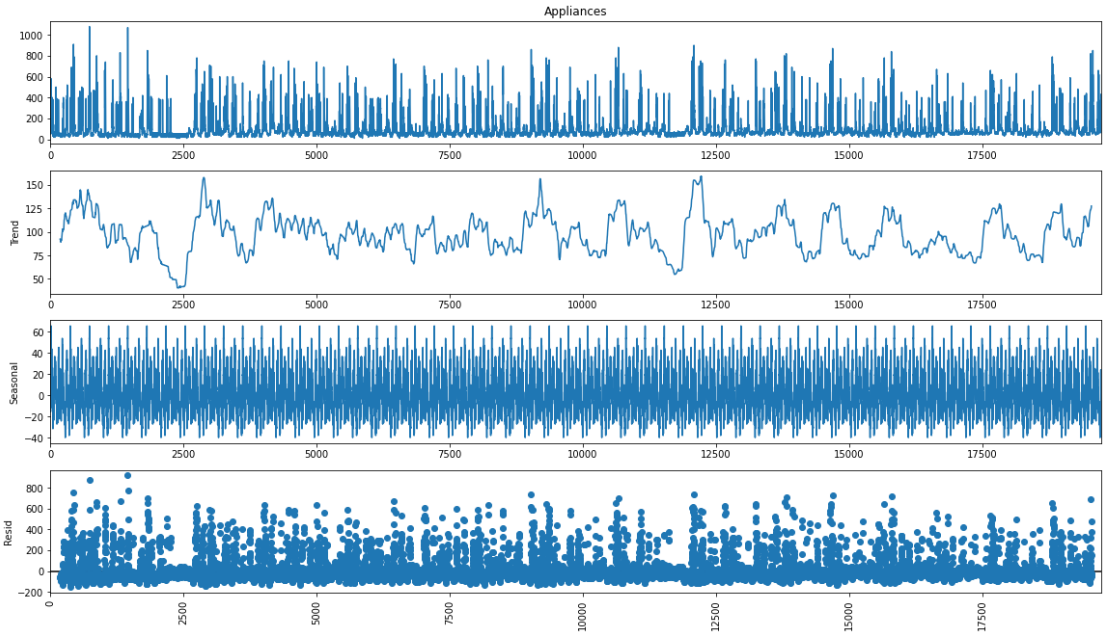

Figure 5

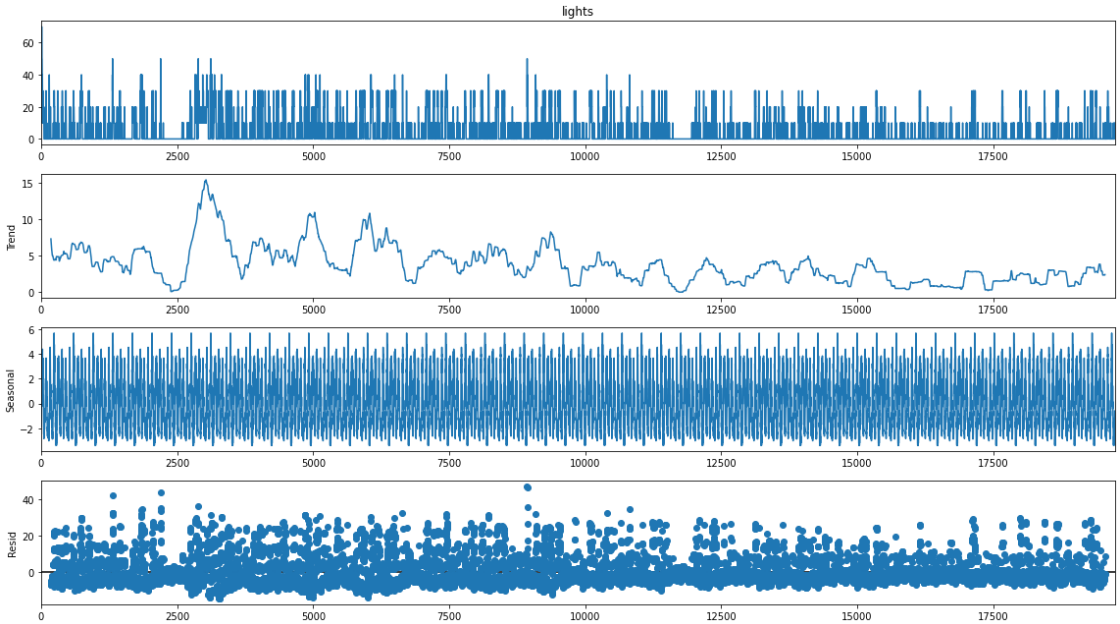

Figure 6

Trends can also have a significant impact on timeseries forecasting. Long-term trends may be indicative of underlying changes in the system being studied, while short-term trends may be influenced by noise or other extraneous factors. Identifying and modeling these trends can help improve the accuracy of the forecast. Our observations of Figures 5 and 6 show that appliances do not follow a certain trend while the lights do show a declining trend over time.

Based on the ADF test for stationarity and the seasonal decomposition of the data, it is clear our data needs transformation. As a result, we differenced our data to stabilize it and reduce the trend and seasonality of it. Now with our data transformed, we could move on to feature selection.

With 28 features in our dataset, there is a very small chance that all of them are statistically significant and need to be included in our model. In order to identify which ones were important for appliances and lights, the Granger Casuality test[6] was applied:

$$y_i = \alpha_0 + \sum_{j=1}^{m} \alpha_j \, y_{i-j} + \sum_{j=1}^{m} \beta_j \, X_{i-j} + \varepsilon_i$$

For the appliances, nine of the 28 columns proved to be statistically significant and will be used in the future model. These variables were temperature in the laundry room (T3) and the kitchen (T1) along with humidity in kitchen (RH_1), living room (RH_2), laundry room (RH_3), ironing room (RH_7) and parent's room (RH_9); lastly, the lights and dewpoint were statistically significant.

For the lights, six columns were statistically significant. These included the temperature in the office room (T4) and the relative humidity in the office room (RH_4), kitchen area (RH_1), living room area (RH_2), ironing room (RH_7) and parents room (RH_9). After learning what features were needed for each model, new data frames were made with only these features so the VAR model could be ran on them.

Hyperparameter Tuning

There are many tests that needed to be initially taken to learn about our data to best prepare it for our model. Once these factors have been evaluated and addressed, a multivariate timeseries forecasting model such as VAR can be developed. The model can also incorporate lags to predict the current value. Now that the appropriate data frames have been made with the significant features, I ran them through the VAR model and used a model function that calculates and prints the lag order; conveniently, this model prints out information for each lag (Akaike information criterion (AIC), Bayesian information criterion (BIC), etc.) and then notes which lag is the best for the model. After applying this to appliances and lights dataframe; I used the AIC value to identify the best lag value. As a result, the lag value chosen for appliances was eight and the lag value chosen for lights was 19.

In summary, developing an effective multivariate timeseries forecasting model requires careful consideration of the stationarity of points, seasonality, trends, correlation and other underlying patterns in our data. By accounting for these factors and incorporating multiple variables into the model, it is possible to create a powerful tool for predicting future outcomes.

Results and Discussion

After the use of the VAR model, we are able to have a better picture of what is occurring with the appliances data. Figure 7 and 8 show the plot diagnostics for appliances and lights. Upon first glance, both residual plots have a relatively even distribution of points around the center value of 0. This is an indicator that the model does not display any obvious seasonality. When looking at the other parts of our plot diagnostic though, it appears there are some notable behaviors.

First, we see the KDE for both Figure 7 and Figure 8 does not closely follow the N(0,1) line which indicates a normal distribution. What this indicates is our residuals may not be normally distributed. This idea is reinforced by the Normal Q-Q in the bottom left corner too. Our data skews away from the line toward the end and the beginning; while this is not necessarily uncommon, the extent that ours does though is. Additionally, the plot does not directly follow the line and we see some fluctuation in the middle. This indicates some skewedness in our data.

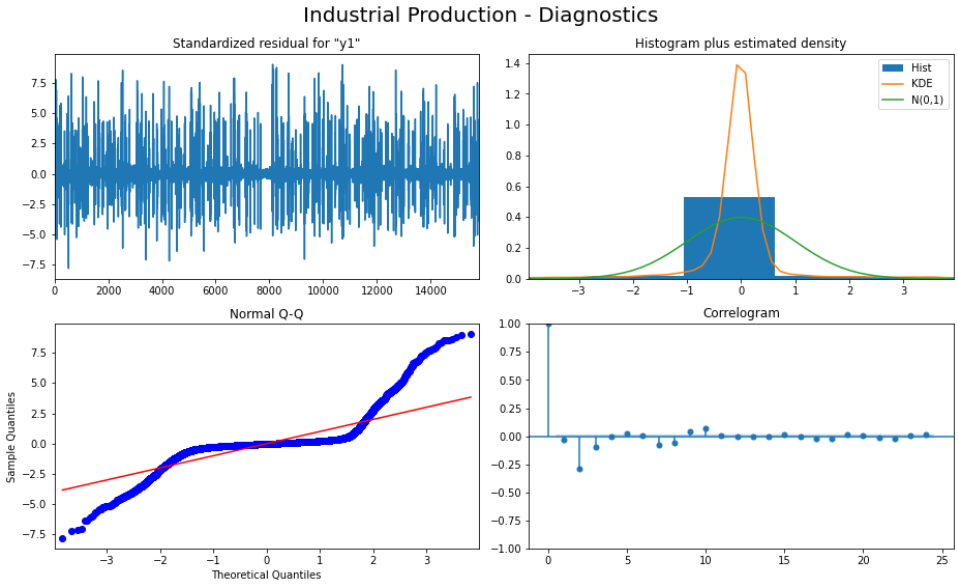

Figure 7

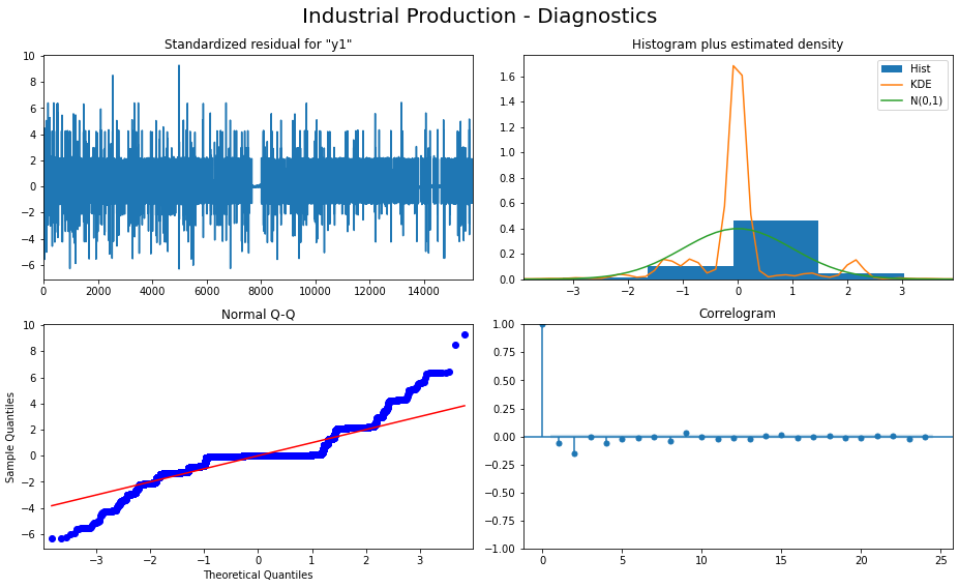

Figure 8

In our VAR model, the predictions were based on the past values of the variables in the model due to the different lag values applied in the appliances model and lights model. What we now see in our final result is in Figure 9, the appliance the predicted values tend to be a bit higher than the actual values. On the other hand, in Figure 10 we see the lights predicted values tend to be lower than the actual. This is should not be surprising; we saw in the plot diagnostics that our data was not following the distribution as we would expect and therefore indicating our model was not satisfactory based on this iformation. We see this reflected in our predictions in Figure 9 and 10.

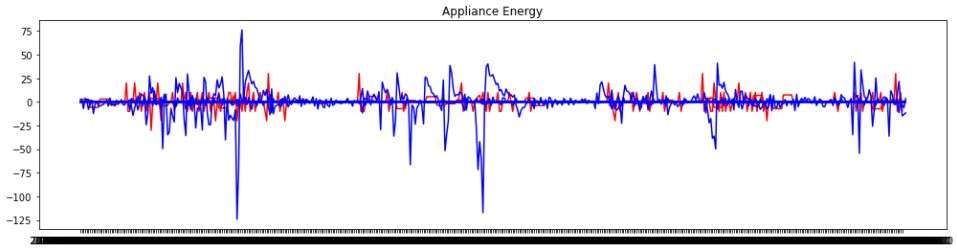

Figure 9

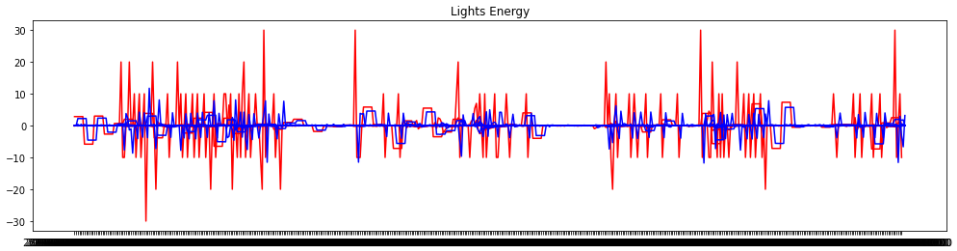

Figure 10

There are a few reasons why the model is not satisfactory for our data. One reason is that there is extraneous information that is not captured by our data set. This could be corrected by finding more data from this study or a similar one and including it in additional research. Another reason is the lag values are not the correct values. While we used the ADF test to identify the best lag value for appliances and lights, there are multiple properties of the lag values that can be used to select the best one. In this study, I used AIC to select the lag value. There were other values that could have been used, including BIC, FPE, HQIC.

Lastly, the variables that were selected could not be the best combination used for appliance and lights. The test used to identify what variables should be used, Granger Causality, usually requires a normal distribution to be followed by the data[7]. As we saw in our plot diagnostics, our data did not follow a normal distribution after the VAR model was applied which could indicate our data was not following the distribution well before. While there are many features in this dataset, this is something that should be considered for future work.

Limitations of Study

There are certainly some limitations when it comes to our study on energy output of appliances and lights. One of the primary limitations is the study's generalization of appliances and lights, which does not break down energy usage for each specific type. This could be a limiting factor, as even the same types of appliances can have different energy outputs, as can the same be said for lights. For example, two refrigerators may have vastly different energy usage patterns depending on their size, age, and other features.

Furthermore, it is important to note that some appliances may not be stored in the same places or locations as others. This could impact their energy consumption, as environmental factors such as temperature, humidity, and light exposure could affect how much energy they use. Additionally, the study's focus on a specific location means that its findings may not be entirely representative of other areas or regions with different climates and environmental conditions. For instance, appliances in areas with colder temperatures may use more energy to maintain their internal temperatures than those in warmer climates.

Another limitation to consider is that there are likely other factors that could impact appliance and light energy output that are not addressed in this study. For instance, the materials used to construct these products, the age of the object, and others could all play a role in how much energy they use. Failure to account for these variables could result in inaccurate predictions of energy usage or energy savings.

Overall, while this study provides valuable insights into the energy usage patterns of appliances in a specific location, it is important to keep in mind its limitations when interpreting its findings. These limitations include the generalization of appliances, the variability of energy usage within the same appliance types, the different storage locations of appliances, the varying environmental conditions, and the other factors that could impact energy output that are not accounted for. By acknowledging these limitations, we can work towards developing more accurate and comprehensive models of appliance energy usage that can be applied across different locations and contexts.

Suggestions for Future Research

First, this dataset has quite a few different seasonalities within it. There are daily seasonalities as the data progresses from day to night and monthly patterns as the weather seasons occur. These could be studied more closely and with other data that has a similar seasonality.

Future research on appliances and their energy output could focus on addressing the limitations of previous studies, such as those mentioned earlier. Specifically, future research could investigate the energy consumption of different types of appliances separately, rather than generalizing across all appliances. This could provide a more nuanced understanding of energy use and help identify areas where energy efficiency improvements can be made.

Moreover, researchers could also study the variation in energy output between different models of the same type of appliance. This could be done by measuring energy consumption for a range of models and analyzing the differences in energy output. This information could then be used to inform consumers and manufacturers about which models are more energy-efficient and potentially spur innovation in energy-efficient design.

Additionally, future research could explore how the location and environment of appliances affect their energy consumption. Researchers could conduct studies in different geographic locations to investigate how outdoor temperature, humidity, and other factors affect energy consumption. This could provide insight into how appliances can be optimized for different regions and climates.

Furthermore, researchers could investigate other factors that affect energy output, such as the materials used to manufacture appliances, the age of appliances, and the air circulation within the appliance. By examining the impact of these factors, researchers could identify additional areas for energy efficiency improvements and potentially guide the development of new technologies that are more energy-efficient.

Overall, there is a significant opportunity for future research in the area of appliance energy consumption, and addressing the limitations of past studies could lead to valuable insights that can inform consumers, manufacturers, and policymakers about how to reduce energy consumption and save money on energy bills.

Reference

1. Bringé, Alison. “Council Post: The State of Sustainability in the Fashion Industry (and What It Means for Brands).” Forbes, Forbes Magazine, 3 Jan. 2023, https://www.forbes.com/sites/forbescommunicationscouncil/2023/01/02/the-state-of-sustainability-in-the-fashion-industry-and-what-it-means-for-brands/?sh=220767a31c82
2. Bullard, Nathaniel. “Automakers Are Investing Billions of Dollars in EVS.” Bloomberg.com, Bloomberg, 5 Aug. 2021,

https://www.bloomberg.com/news/articles/2021-08-05/automakers-are-investing-billions-of-dollars-in-evs?leadSource=uverify+wall.
3. Benjamin, Heather. "LEED-Certified Office Buildings Found to Bring High Sale Premiums." U.S. Green Building Council, U.S. Green Building Council, 27 Jan. 2022, https://www.usgbc.org/articles/leed-certified-office-buildings-found-bring-high-sale-premiums.
4. "Vector Autoregressive Models for Multivariate Time Series." *SpringerLink*, Springer New York, 1 Jan. 1970, https://link.springer.com/chapter/10.1007/978-0-387-32348-0_11.
5. Menegaki, Angeliki. "A Guide to Econometrics Methods for the Energy-Growth Nexus." Dickey-Fuller Test - an Overview | ScienceDirect Topics, 2021, https://www.sciencedirect.com/topics/economics-econometrics-and-finance/dickey-fuller-test.
6. Rossi, Barbara. "4.1.1 Do Traditional Macroeconomic Time Series Granger-Cause Inflation and Output Growth?" *Granger Causality Test - an Overview | ScienceDirect Topics*, Handbook of Economic Forecasting, 2013, https://www.sciencedirect.com/topics/social-sciences/granger-causality-test.
7. Chvosteková, Martina, et al. "Granger Causality on Forward and Reversed Time Series." *Entropy (Basel, Switzerland)*, U.S. National Library of Medicine, 30 Mar. 2021, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8066447/#:~:text=Besides%20the%20normal%20distribution%2C%20which,of%20predictive%20errors%20are%20considered.

Code Reference

Seabold, Skipper, and Josef Perktold. "Statsmodels: Econometric and statistical modeling with python." Proceedings of the 9th Python in Science Conference. 2010.

Appendix

https://github.com/parksjr5/Energy_Forecasting