APPLIANCES ENERGY PREDICTION

Sara Albalooshi
Data Analytics Fellow,
General Assembly, Bahrain

Abstract:

Energy consumption is on the rise globally, driven by increasing demand, which in turn leads to higher electricity bills. Certain locations worldwide show excessive energy usage in household appliances. This project aims to propose strategies to reduce energy consumption in home appliances by understanding the trend and predicting energy usage using regression model.

In this energy prediction challenge, machine learning will be employed to forecast household appliance energy usage by considering variables like temperature, humidity, and pressure. This report primarily focuses on utilizing linear regression for energy consumption prediction.

1.Problem Statement

The objective is to forecast household appliance energy consumption by analyzing a dataset collected over about 4.5 months, with sensor readings taken every 10 minutes. Monitoring of temperature and humidity in the house was conducted through a ZigBee wireless sensor network, where data transmitted approximately every 3.3 minutes. Weather information from the Chievres Airport station in Belgium was acquired from a public dataset. Energy consumption data was logged every 10 minutes using m-bus energy meters.

1.1 Data Set

Feature	Description
lights	Energy use of light fixtures
T1	Temperature in kitchen area
RH_1	Humidity in kitchen area
T2	Temperature in living room area
RH_2	Humidity in living room area
T3	Temperature in laundry room area
RH_3	Humidity in laundry room area
T4	Temperature in office room
RH_4	Humidity in office room
T5	Temperature in bathroom
RH_5	Humidity in bathroom
T6	Temperature outside the building
RH_6	Humidity outside the building
T7	Temperature in ironing room
RH_7	Humidity in ironing room

Feature	Description
T8	Temperature in teenager room 2
RH_8	Humidity in teenager room 2
T9	Temperature in parents' room
RH_9	Humidity in parents room
T_out	Temperature outside (from Chievres weather station)
Press_mm_hg	Pressure (from Chievres weather station)
RH_out	Humidity outside (from Chievres weather station)
Windspeed	Wind speed (from Chievres weather station)
Visibility	Visibility (from Chievres weather station)
Tdewpoint	Tdewpoint (from Chievres weather station)
rv1	Random variable 1
rv2	Random variable 2
Date	Date and time format
Appliances	Energy used by appliances (Target Feature)

2. Introduction

Being able to compute the consumption patterns of daily home appliances, ranging from washing machines to televisions and microwaves, is crucial as it accounts for a significant portion of a household's energy demand and cost . Analyzing this data can offer valuable insights into optimizing electricity utilization.

This involves predicting appliance electricity usage based on a multitude of influential factors. Environmental conditions near appliances, such as temperature, humidity, light exposure, and vibrations, play pivotal roles in shaping energy consumption.

Developing predictive models for this task holds immense utility across various domains, including anomaly detection in energy consumption and forecasting energy demands.

Employing regression models facilitates this predictive analysis. Given the nonlinear nature of the problem, accurately estimating energy consumption solely based on the provided features poses a challenge due to their interdependencies. Instead of opting for a time-series approach, the project will leverage machine learning techniques to address this intricate task.

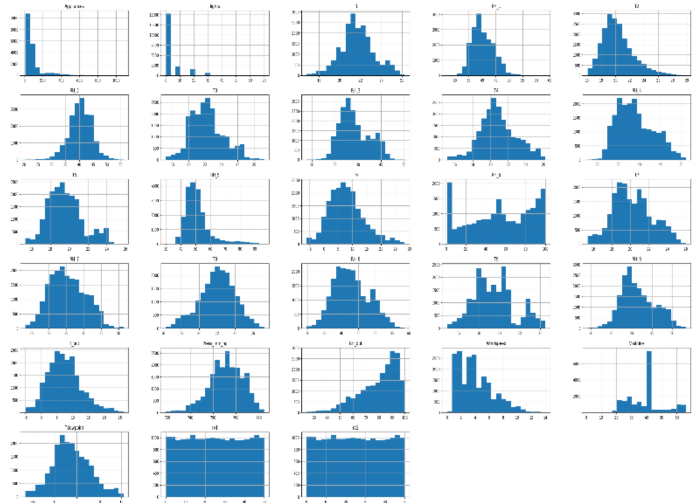
3. Data Exploration

At the outset of my analysis, I delved into a dataset containing 19,735 rows and 29 columns. Upon scrutiny, I found no missing data, ensuring data integrity. Most columns hold numerical values, apart from the 'date' column, which I plan to adjust for better analysis later.
Notably, the 'Appliances' column, representing appliance energy use, spans from 10 to 1080 watts. This range offers valuable insight into energy consumption patterns, laying a solid foundation for deeper exploration.

3. Data Visualization

In this section, I utilized graphical representations to find trends, patterns, and validate assumptions within the dataset. These tools were crucial in gaining insights and understanding the underlying characteristics of the data.

The frequency distribution of all data was visualized as depicted below.

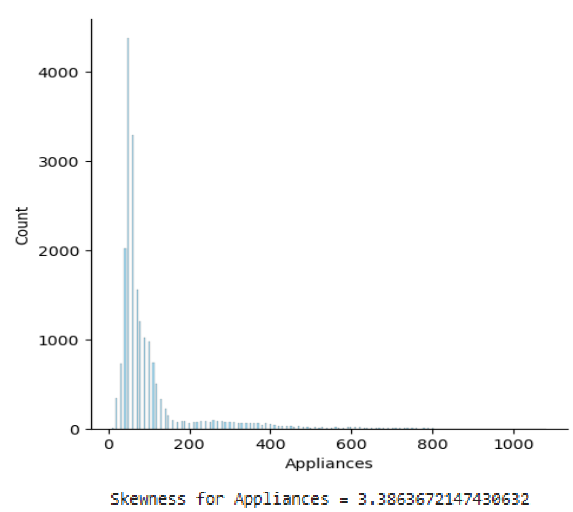


Most temperature readings seem to have normal distribution, apart from T9. The humidity values as well, RH_6 and RH_out, exhibit a normal distribution.

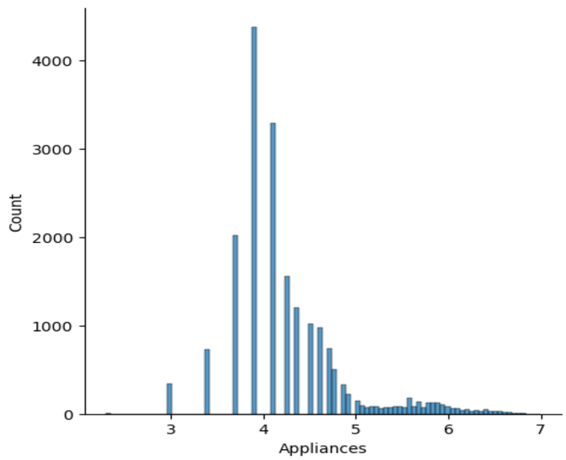
It is also clear that Visibility, Windspeed, and Appliances display skewness. The random variables rv1 and rv2 seem to be consistent values across all recordings.

<u>Appliances Column</u>

To understand the trends within the 'Appliances' column. I employed a 'distplot,'



the plot showed that a significant portion of appliance power consumption falls within the 1-250Wh range. The column displays positive skewness, with a concentration of values around the mean of 100Wh, alongside the presence of outliers.

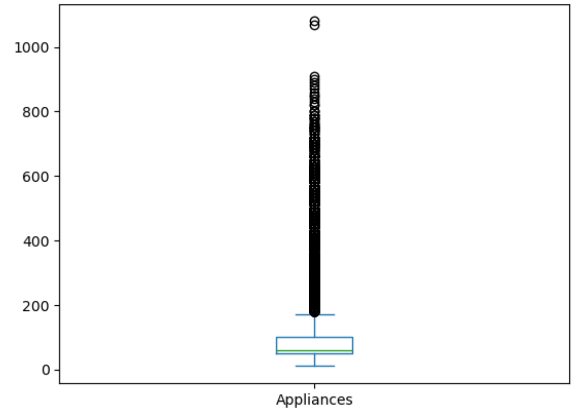


Skewness for Appliances = 1.165816888806544

After implementing a log-transformation on the 'Appliances' data, the skewness reduced to 1.16, as indicated in the chart. This adjustment aims to enhance the data's distribution for more robust model performance.

Outliers handling

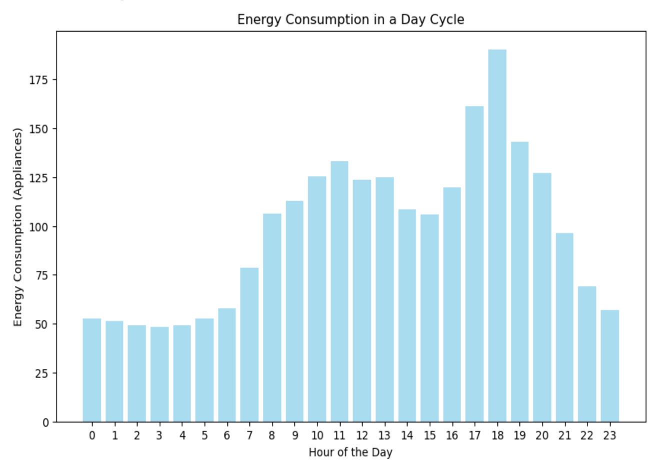
Approximately 90.29% of values in the 'Appliances' column are less than or equal to 200Wh, indicating a prevalent lower energy consumption trend. However, the existence of outliers, particularly with a maximum consumption of 1080Wh, suggests instances of unusually high energy usage. These outliers could potentially impact our machine learning model adversely in subsequent steps.



Given that outliers constitute 9.71% of the data and may introduce bias, it is imperative to address them. Consequently, in the forthcoming stages

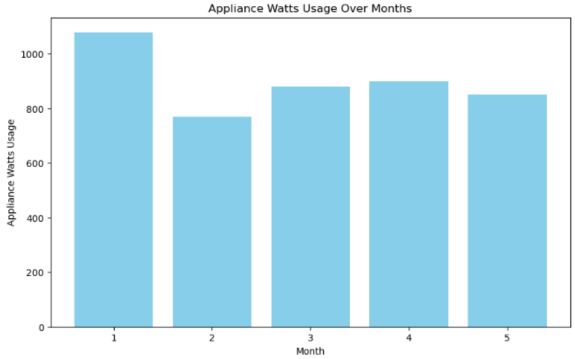
Energy Consumption with time

From the bar chart we can understand the following:



Electricity consumption is highest in the evening hours between 4:00 pm and 8:00 pm. During the night hours, from 11:00 pm to 6:00 am, power load remains below 50 Wh, which is expected as most appliances are either switched off or on standby during this time. Whereas between 9 am and 1 pm, consumption exceeds 100 Wh, likely due to activities such as breakfast. Subsequently, consumption decreases to below 100 Wh.

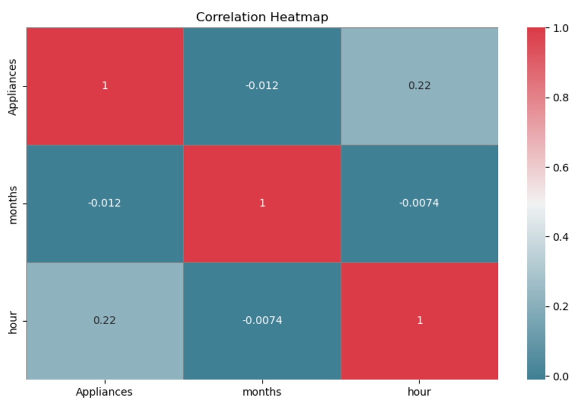
Next, I initially employed a bar plot to identify the month with the highest energy consumption.



However, as this method did not provide a clear answer, I resorted to calculating the total energy consumption for each month, leading to the following findings:

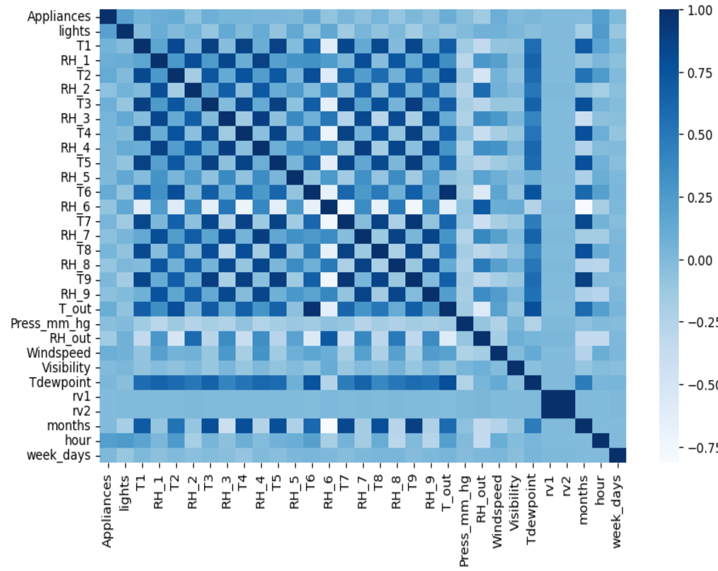
```
months
3     432800
4     427200
2     421550
5     362950
1     283510
```

Correlation with Time features



After conducting a correlation analysis, it is apparent that the 'Appliances' (target) feature demonstrates the strongest correlation with the hour variable. Interestingly, there seems to be no notable relationship between 'Appliances' and the months.

Correlation with numerical features



Based on the plot all the temperatures showcase positive correlations with the target variable 'Appliances.' Particularly, T9 exhibits a high correlation with T7 (0.942), while T_out is strongly correlated with T6 (0.973). This indicates a robust relationship among temperature metrics, illustrating how changes in outdoor temperature can affect indoor spaces.

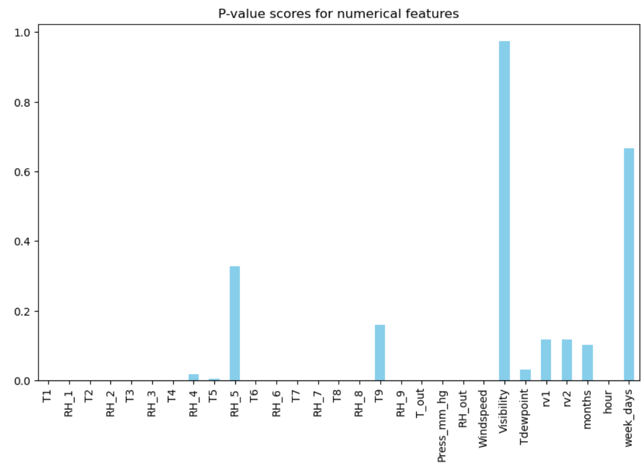
Moreover, I observed an inverse correlation between RH_6 which is an (outside humidity) and all temperature features, due to the impact of humidity on the temperature.

As air temperature rises, the air's capacity to retain moisture increases, resulting in a decrease in relative humidity.

4. Feature Selection

For feature selection the below method was used:

F-regression- I utilized the F-regression method for feature selection due to its ability to statistically assess the importance of features relative to the target variable. This technique helped to identify key features, rank them by significance. As a result, I eliminated redundant features such as RH_4, RH_5, months, Tdewpoint, T9, Visibility, rv1, rv2, and weekdays.



Additionally, I removed the date, dates, and lights columns as they did not contribute significantly to the analysis**.**

```
data.drop(["date","dates","lights"],axis=1,inplace=True)
```

5. Feature Engineering

As previously mentioned, I had detected outliers in the 'Appliances' column, representing around 9% of the dataset and impacting my machine learning models negatively. To tackle this challenge, I employed the IQR (Interquartile Range) method for outlier detection and removal. By utilizing this technique, I aimed to cleanse the dataset of outlier-related anomalies, potentially boosting the precision.

Finalized Features

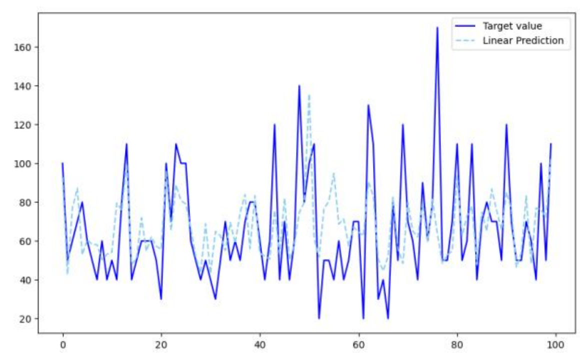
After evaluating the outcomes of removing the uninformative columns, I opted to proceed with these columns as my features for further analysis.

```
final_feature = ['T1', 'RH_1', 'T2', 'RH_2', 'T3', 'RH_3', 'T5', 'RH_6', 'T7', 'T8', 'RH_8', 'Press_mm_hg', 'Windspeed', 'hour']
```

6. Making Models

I used Linear Regression processes to train my model

Note- The lower the value of the RMSE and the higher the R_2 the better the model.



7. Conclusion

Initially, the data was identified as time series due to its regular time intervals, but the decision was made not to implement time series techniques due to a lack of expertise in the area. Subsequently, matplotlib and seaborn were used for Exploratory Data Analysis, creating various plots such as scatter, bar, boxplot, and heat map. Notably, energy consumption peaks in March and dips in January, with temperature fluctuations impacting power usage. Humidity inversely affects energy consumption, and the hour of the day significantly influences it. Evening hours exhibit high electricity consumption, especially on weekends. During feature engineering, outliers were removed from the model. Ultimately, the model contained crucial features for prediction, emphasizing the significance of the hour feature.

Recommendations

1. Gain more insights by examining occupant demographics, building features, and other factors to improve energy consumption predictions.

2. Gather more data to overcome the challenge of short data analysis periods to capture seasonal energy variations, gaining a thorough understanding of energy trends over the year.

3. Increasing sensor numbers and accuracy can enhance prediction capabilities.

References-

- Stack overflow
- Medium
- GeeksforGeeks
- Kaggle
- Analytics Vidhya