Project 1: Celebaltech: Non interfering load monitoring.

Vaishnavi Shivkumar

NILM (Non-Intrusive Load Monitoring): NILM, also known as energy disaggregation, is a process to estimate the energy consumed by individual appliances given the whole house's energy signal. NILM techniques can be very beneficial for energy conservation because they can help identify which appliances consume the most power, at what times, and suggest ways to reduce consumption.
Our data set is in reference to http://arxiv.org/abs/1004.0456.
http://archive.ics.uci.edu/dataset/235/individual+household+electric+power+consumption

Method 1:

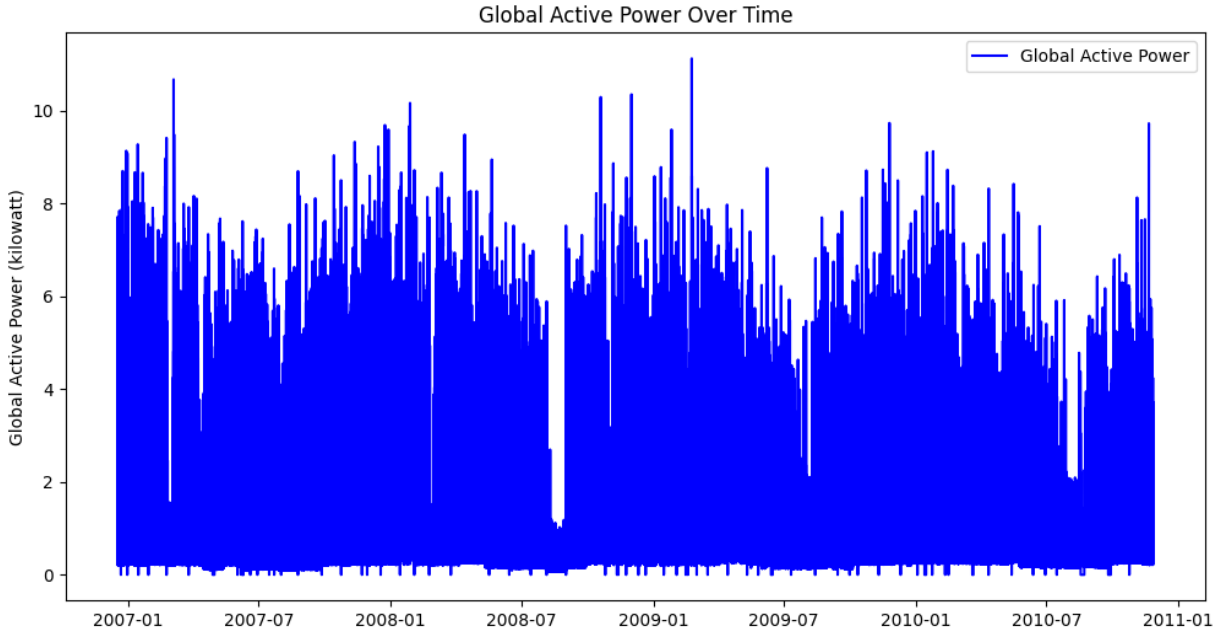
We attempt to find the presence of a signature called reference but comparing its fast fourier transformation with a reference signal's fft.
FFT (Fast Fourier Transform): FFT is an algorithm used to compute the discrete Fourier transform (DFT) and its inverse. The Fourier Transform is a method that transforms one complex-valued function of a real variable into another. In the context of signal processing, it provides the frequency components of the original signal. It can reveal important characteristics of a signal that are not visible in the time domain.

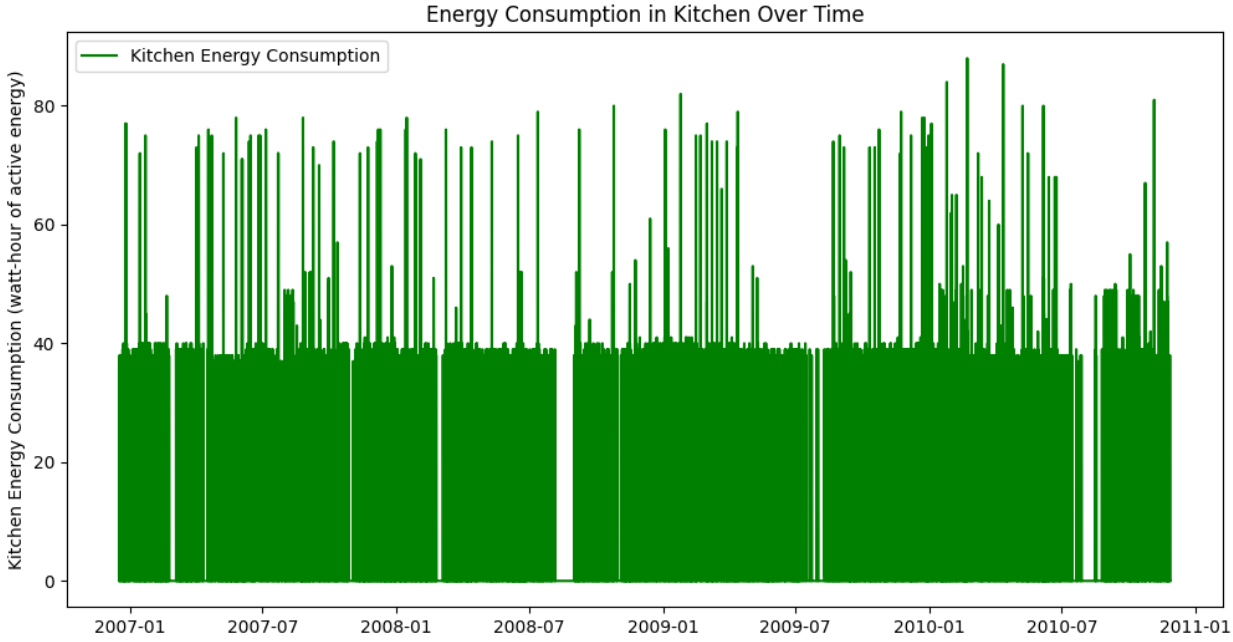
I attempted to heuristically set thresholds to analyze this particular databset. This dataset has global active power, Global reactive Power, Voltage, Global intensity, sub_metrix_1 (which represents the power usage in the kitchen); sub_metric_2 (which represents the power useage in the laundry); sub_metric_3 (which represents the power usage by a few heavy appliances, like water heater and air conditioner.) This dataset DOES NOT contain meta data of individual appliance. This poses as a huge downside for this data set. Regardless, i have modelled the classification problem differently. Here, I have defined a new column, which is called total_power, which is the (kitchen + laundry + heavy appliances) power.

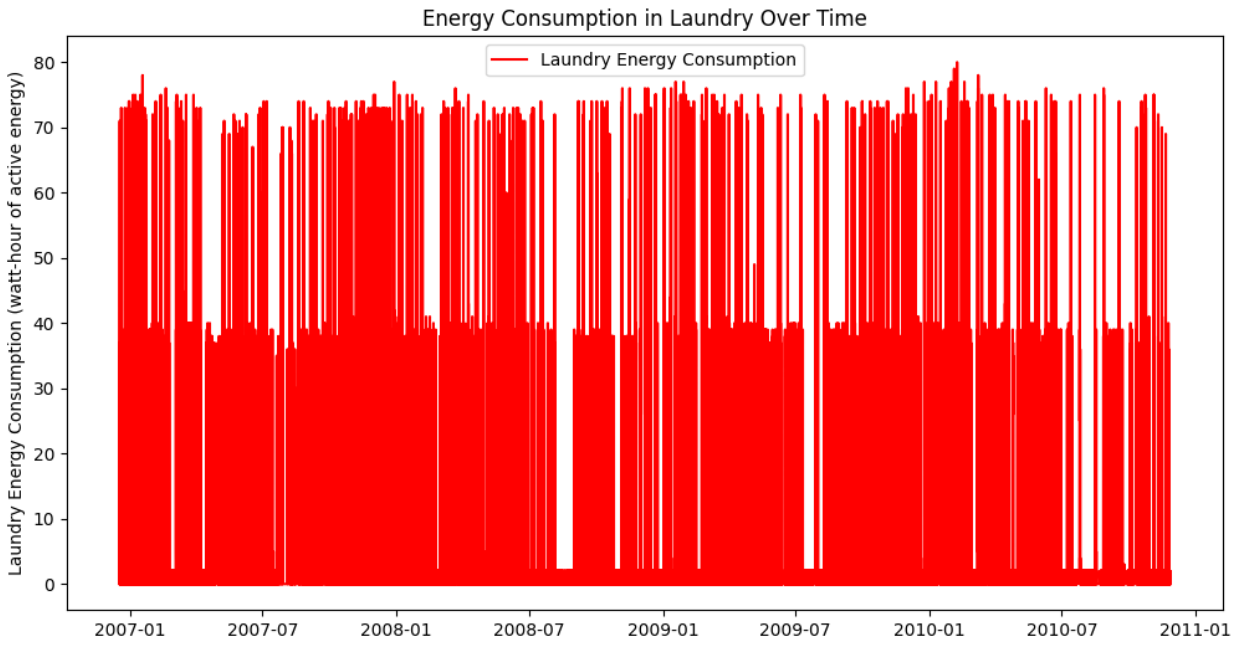
We take the fourier transformation of the reference signal (our appliance) and compare it to sliding windows of the frequency components of multiple parameters. There are some nan values in the dataset, which we replaced with zero.

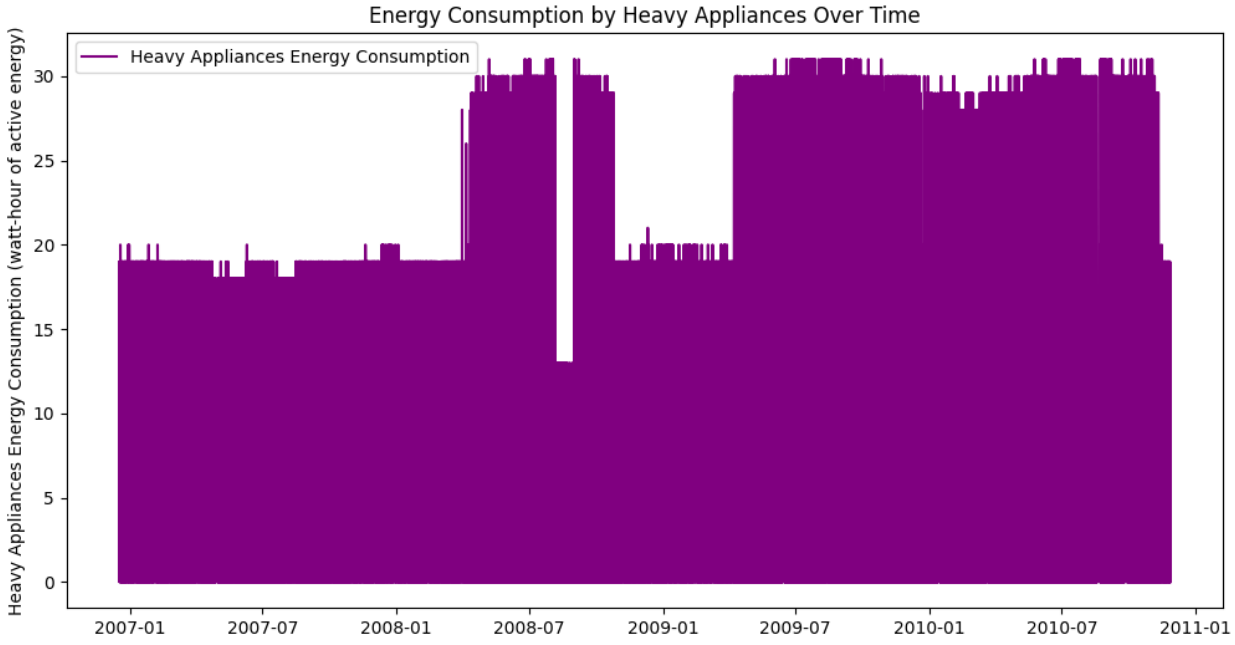
We started our analysis by plotting the graphs (kilowatt vs time) for global active power, energy consumption in kitchen, laundry and by heavy appliances. We took a measure called other appliances which corresponds to other appliances in the house using power. This is to

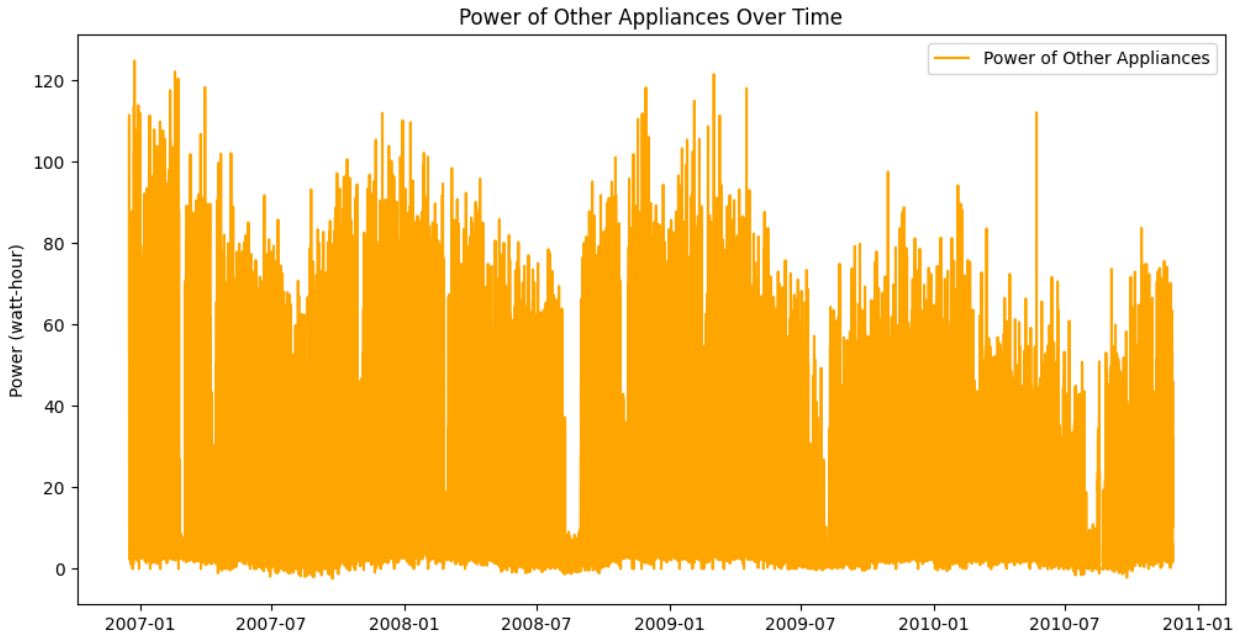
understand how our graphs look like. We also took a subsection of Kichen power, and plotted it in order to heuristically find ourselves a reference.





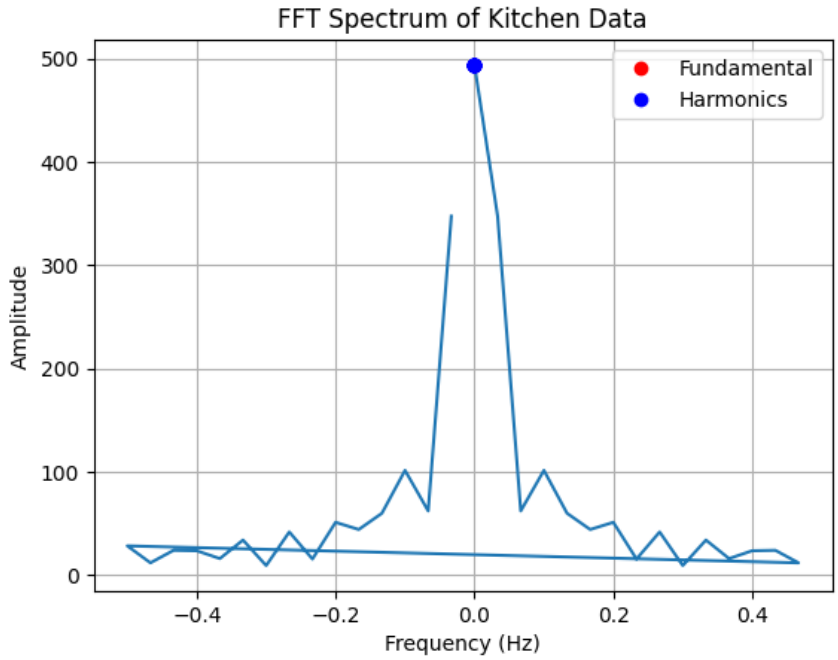

Energy Consumption in Laundry Over Time
Laundry Energy Consumption
Laundry Energy Consumption (watt-hour of active energy)
0
10
20
30
40
50
60
70
80
2007-01
2007-07
2008-01
2008-07
2009-01
2009-07
2010-01
2010-07
2011-01

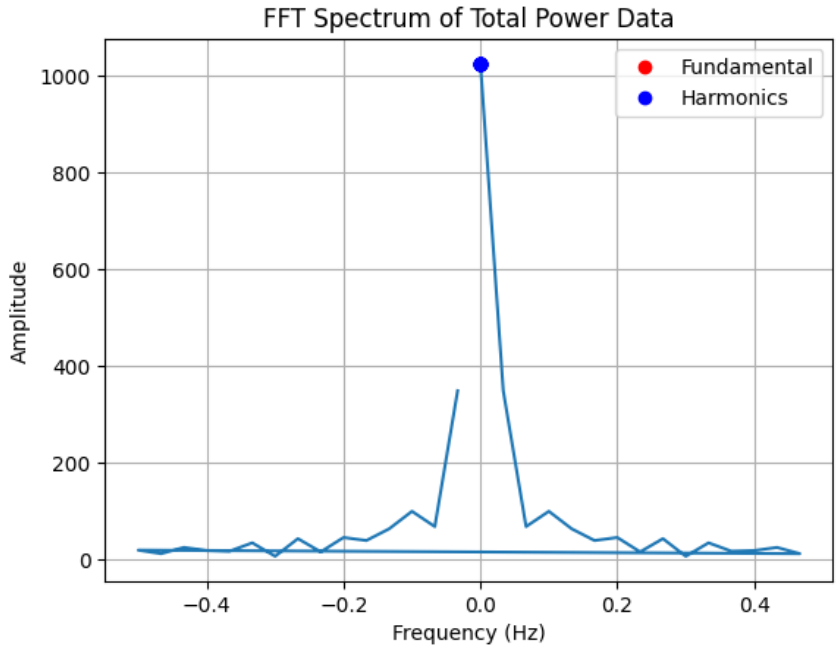

Energy Consumption by Heavy Appliances Over Time
Heavy Appliances Energy Consumption
Heavy Appliances Energy Consumption (watt-hour of active energy)
0
5
10
15
20
25
30
2007-01
2007-07
2008-01
2008-07
2009-01
2009-07
2010-01
2010-07
2011-01



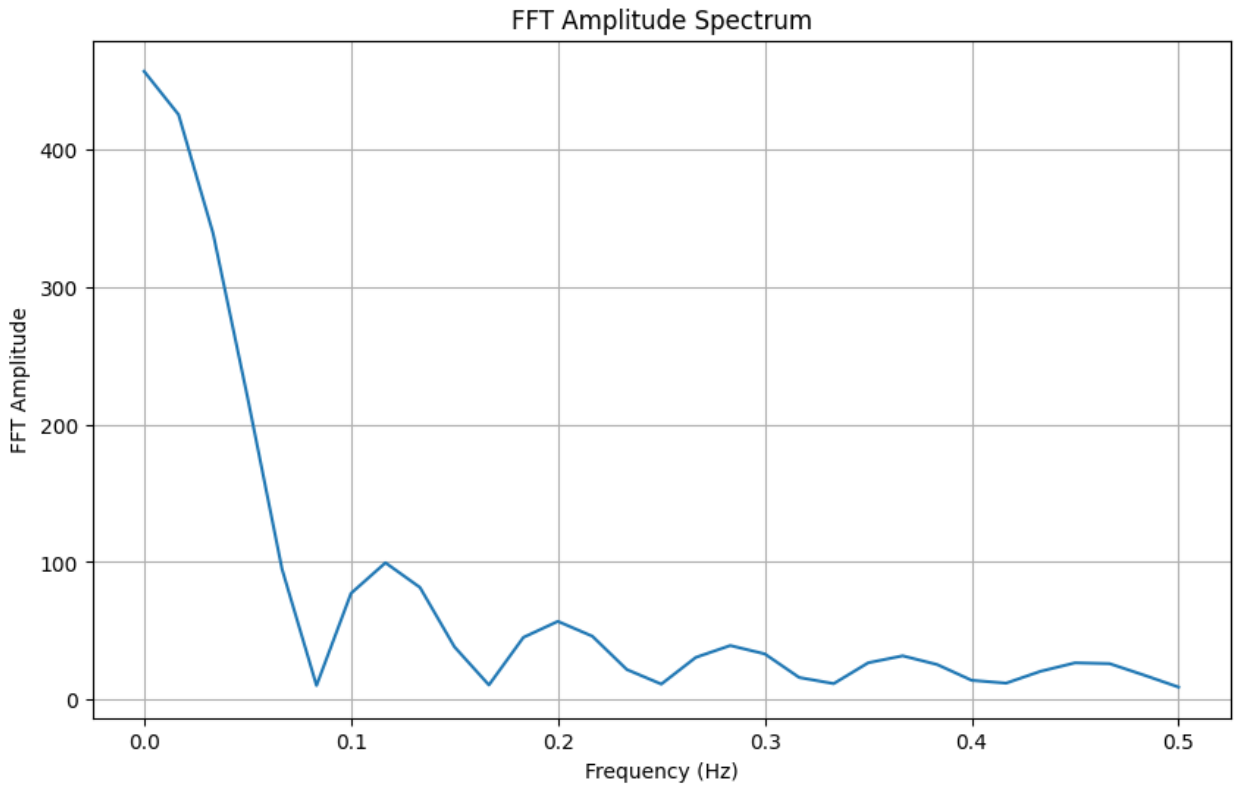
We are assuming the above part of our signal to be our reference (between index 1021 and 1081) . We do see a slight repetition, but that hardly changes things when we are comparing sliding windows of fast fourier transforms.

The next part of our analysis is taking the Fast fourier transformation of a fixed window of total power and kitchen. This is just to analyze how the FFT plot looks like.





We now plot the amplitude spectrum and the corresponding frequencies of the reference signature. Unlike the previous situation, we model real like by taking the absolute value of frequency components.



Now, when we traverse through the signal in time domain (using a sliding window of pane_size = 60); we slide it, while increasing start index by 5minutes everytime. Essentially sliding the pane first between 0-60, then 5-65 etc. This 5 is called the increment that can be a changable value. Here we check for the similarly of our reference fast fourier transformation and the fast fourier transformation.

In the analysis, one of the primary objectives is to identify when the spectral power (frequency count) of the 'total_power' signal can be expressed as a linear combination of the spectra of different appliance signals (kitchen, laundry, heavy appliances, etc.). This is represented by the equation ax1 + bx2 + cx3, where x1, x2, and x3 correspond to the frequency counts of kitchen, laundry and heavy appliances, respecitvely . A minor condition applied is that the spectral power of 'total_power' should exceed that of 'kitchen' at least 75% of the time.

We take a sliding window of increment 5 minutes, where analysis takes place comparing the reference to 0-60, 5-65 etc. The increment here is 5, and can be changed. Let us take a particular sliding window.
The analysis also includes some conditions for data classification based on the presence or absence of a 'reference frequency' in the 'total_power' and 'kitchen' signals. This is done by calculating how closely the spectra of the total_power and kitchen signals resemble the spectral of the reference window. This resembelence is measured as a Euclidean distance between the spectra, with lower values indication a closer resembelence. If both kitchen and total_power have distance more than this threshold, they are assigned a value 0. If both have distance less than threshold, the sliding window is assigned a 3. If total power distance exceeds the threshold

and kitchen doesnt, it essentially means that this reference is present in total power, but it doesnt correspond to kitchen, It rather corresponds to some other frequency components (perhaps from laundry or heavy appliances) this gets assigned a value 1 (these are collission errors). If total doesnt match but kitchen matches, it is assigned a 2 (this is the primary spot for errors which doesnt come from collisions)

The accuracy depends on the values assigned a 0 and a 3. We ensure that total_power exceeds the kitchen, and the rest are discarded (this policy should be true astleast 75% of the time which is represented as a variable called mean.)
If a window of data assigned 0 or 3 doesn't meet these conditions, it is 'discarded' and pushed into a separate dataframe for further analysis. This strategy is an effective way to reuse the data that initially failed to meet the criteria.

In the follow-up analysis of the discarded data, the conditions are relaxed by increasing the thresholds min_total_power_dist and min_kitchen_dist by 25. This essentially doesnt do much, since already discarded data fullfill the initial threshold. This just tightens conditions to return/exit our recursive loop.

However, this is counterbalanced by relaxation of another condition: the requirement for 'total_power' to exceed 'kitchen' is now enforced at least 60% of the time, compared to the initial 75%. To prevent infinite recursion and progressively worse data, the code ensures that this threshold never falls below 40%.

Throughout the entire process, all the classification outcomes are stored in a results dictionary. This includes outcomes from multiple recursions, which helps in tracking how the classification results evolve as the conditions are modified across iterations.

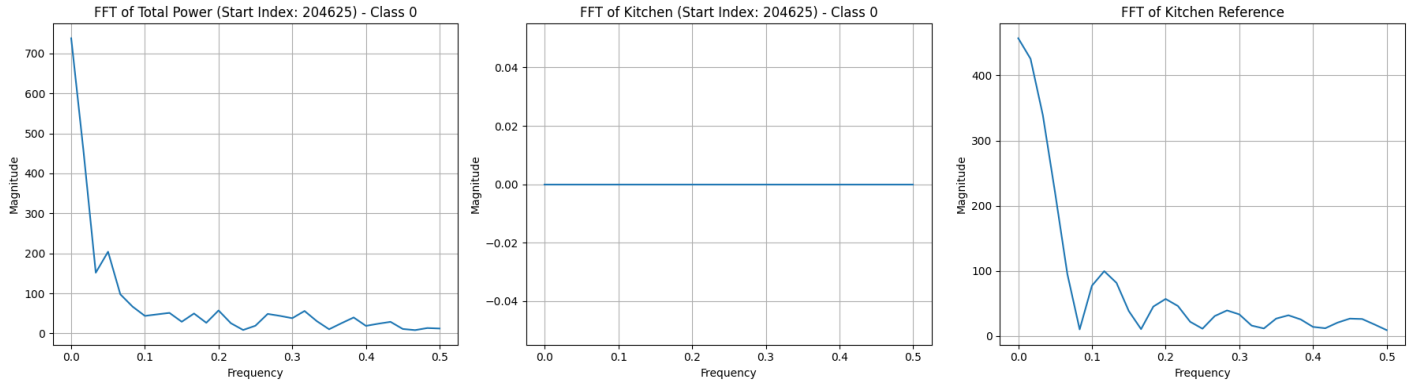
This is the output when we take the total_power minimum distance threshold as 50, the kitchen minimum distance threshold as 50 (the difference inamplitude (here coefficients corresponding to the frequency components) of total_power and reference and kitchen and reference respectively should not exceed 50); We took the pane_size as 60 (sliding window pane of reference considers 60 minutes of data) We take the increment of 15 minutes, which means that the sliding window start point moves up by 15 minutes.

According to the previous values of input, we find that the accuracy is 77.2649914906397%

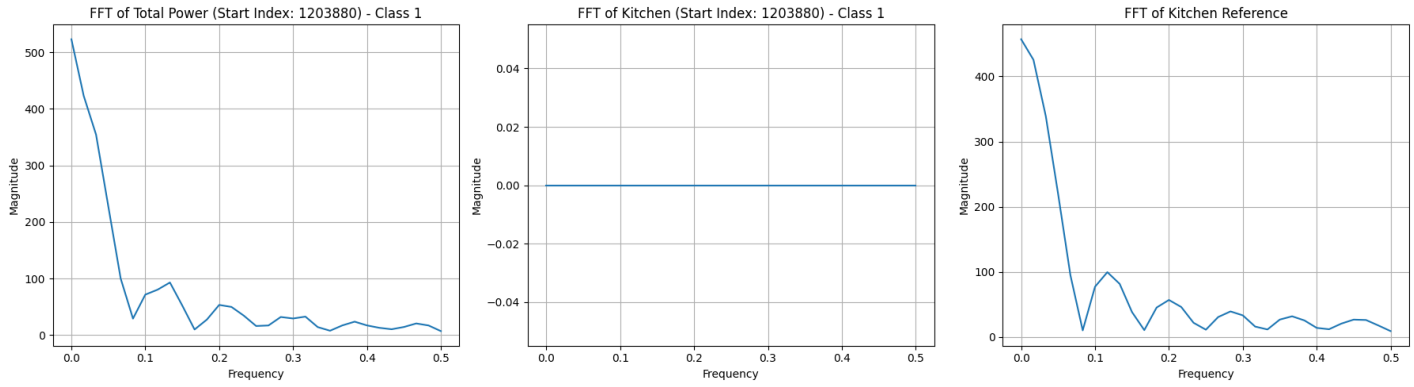
We can see here that values 0 and 3 correspond to when our algorithm is accurate. Index 1 is positions of collissions (or where some other room has similar frequency response as the referenced appliance so it acts as a collission error) Index 2 is the position where the collisions cancel out each other. , and is seeingly minimum.

For value 3: when the reference appliance is in both kitchen and total power. Our reference is being used in the kitchen.

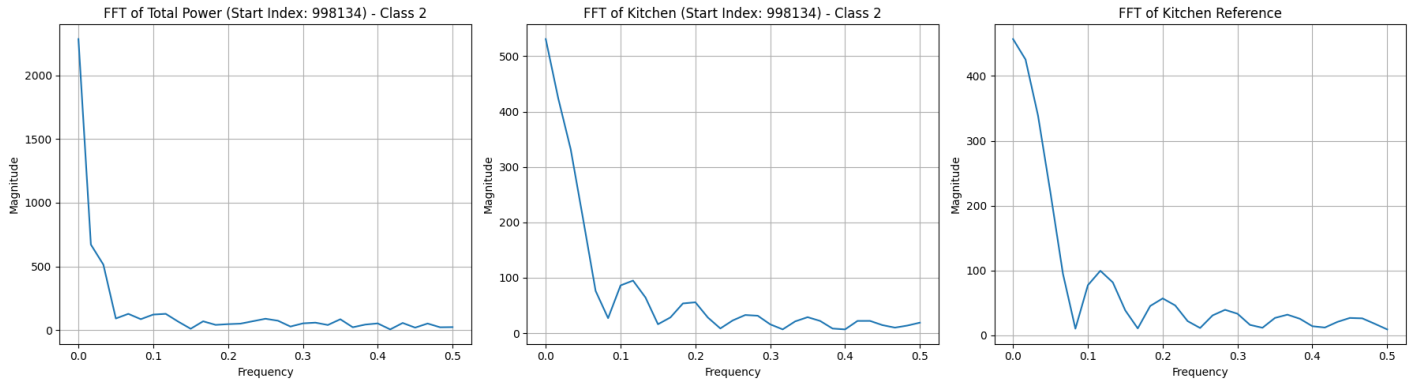
For value 0: our reference is not in both kitchen and total_power, hence isnt in use.



For value 1: corresponding to our reference being shown in total power but not in the kitchen. Hence we can say that the reference appliance is not being used in the kitchen. It or something similar to it is being used in other rooms, and hence is reflecting in our total power.



For value 2: s us that the reference is being used in the kitchen, but it isnt reflection in total power. This is an observation where the collisions cancel out eachother.



In conclusion, our algorithm here is called accurate if it correctly identifies if our appliance called "reference" is being used in the kitchen AND is reflected in total_power when its being used.
So it is accurate if:
1) It is not being used, and hence doesnt show up in kitchen or total_power.
2) It is being used and shows up in both kitchen and power.

Our algorithm additionally tells us if an appliance is being used, which is similar to reference in other rooms/domains within the house.
Further things we can do is handle collission error (values 1 and 2) by comparing more parameters and by employing more rigourous algorithms that accurately change our threshold parameters, rather than heuristically (the method that has been employed right now)

Method 2:

This report presents an analysis of the household power consumption data using wavelet transform and cross-correlation techniques. The goal is to identify similarities between the reference pattern and the 'Kitchen' and 'Total Power' time series. The key findings and inferences from the analysis are summarized below:

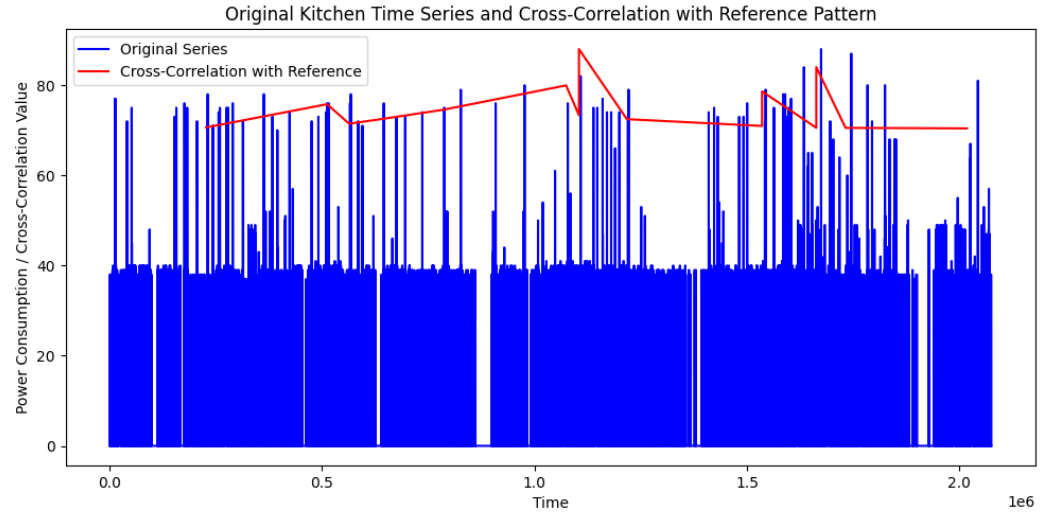
Data Preprocessing: The data is preprocessed by cleaning missing values and computing the 'total_power' variable as the sum of 'Sub_metering_1', 'Sub_metering_2', and 'Sub_metering_3'; which are respectively renamed as Kitchen, Laundry and Heavy Appliances.

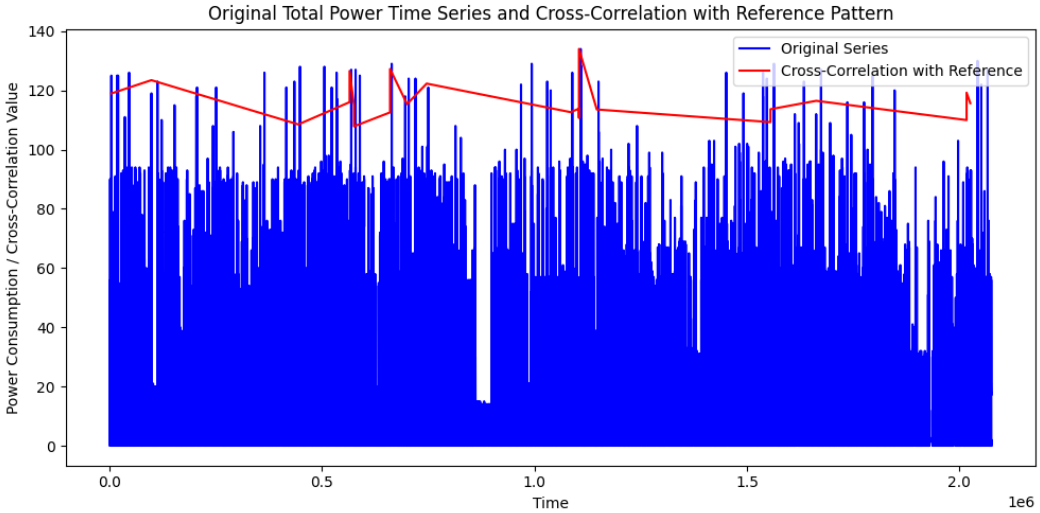
Wavelet Transform: The reference pattern, extracted from the 'Kitchen' time series (and is assumed to be an appliance), is transformed using the Haar wavelet. The wavelet transform decomposes the data into different scales, representing different levels of detail. This provides a way to analyze the reference pattern at different resolutions.

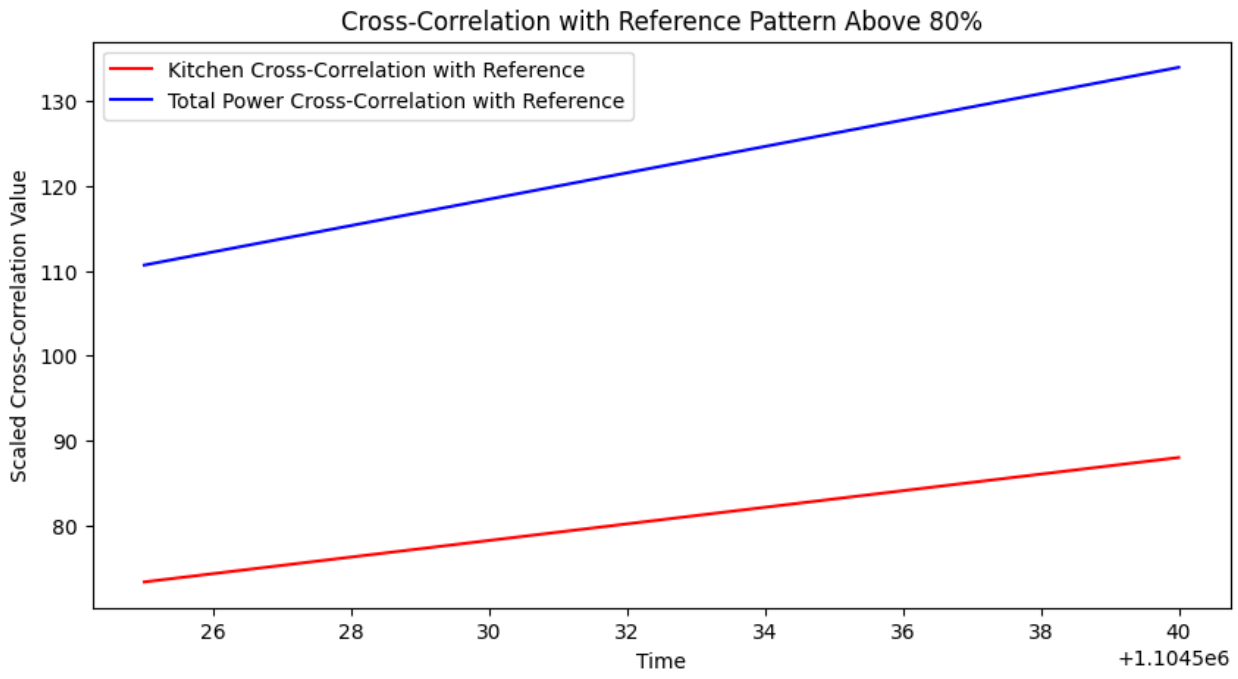
Cross-Correlation Analysis: The wavelet coefficients of the reference pattern are compared with the coefficients of the 'Kitchen' and 'Total Power' time series at each scale. The cross-correlation measures the similarity between the reference pattern and the time series at each point in time. The correlation is performed for each scale and the results are combined to obtain a measure of similarity.

Sliding Window Approach: To identify points of high similarity, a sliding window approach is used. A window of the reference pattern size is slid along the 'Kitchen' and 'Total Power' time series with a step size of 15 indices. The wavelet transform is applied to each window, and the cross-correlation values are computed. This process results in time series that represent the similarity between the reference pattern and the 'Kitchen' and 'Total Power' time series.

Scaling and Visualization: To compare the cross-correlation values with the original time series, scaling factors are calculated based on the maximum values of the cross-correlation series. The cross-correlation values above 80% of the maximum are highlighted in the plots to indicate significant correlations.







Observation and Inference: The analysis reveals a positive slope in the cross-correlation values for both the 'Kitchen' and 'Total Power' time series. This suggests an increasing alignment of the usage patterns of both areas with the reference pattern over time. The inference is that there is a consistent and growing similarity between the usage patterns of the kitchen and the overall household compared to the reference pattern.

It's important to note that correlation does not imply causation, and further investigation is required to understand the underlying factors driving these patterns. The positive slope and similarity in the cross-correlation values provide valuable insights into the alignment of usage patterns but should be interpreted in conjunction with other information.

This analysis demonstrates the potential of wavelet transform and cross-correlation techniques in understanding the similarity between power consumption patterns. It provides a foundation for further research and exploration in the field of non-intrusive load monitoring (NILM) and energy disaggregation.