Forecasting the Electric Power Consumption for a House

Week 1 (Data Series Visualization)

Bipul Bishwas, Ekhtear Khan, Chantelle Amoako-Atta

1 INTRODUCTION

Background and motivation

According to [1], electric power consumption of households is an important factor for energy planning and management, as well as for understanding the environmental impact of electricity use. In this report, we aim to analyze and forecast the electric power consumption given data from one household with a one-minute sampling rate over a period of almost four years (between December 2006 and November 2010). The data set we use is the Individual household electric power consumption Data Set from the UCI Machine Learning Repository [2] . It contains 2,075,259 measurements of different electrical quantities and some sub-metering values that indicate the energy consumption of specific appliances. The data set also contains some missing values (nearly 1.25% of the rows), which we will visualize and handle in the data preprocessing step. The main objective for this week is to get **to know the data, the sampling rates, if there are missing values and how much of the data we need**. We did the following to achieve this.

- Checked for and handled missing values.
- Visualized and made assumptions about the variables in the data using some time series plots like time series box plots and line plots.
- Visualized and studied the trends, seasonalities and residuals of some selected variables in the data over varying sampling rates (including daily, monthly, and yearly rates)

We used Python to perform the data analysis and our codes can be accessed here.

We were given that the `total_energy_consumption` is the variable to be forecasted is calculated using the formula in equation (1) below.

$$y = global\_active\_power * \frac{1000}{60} - sub\_metering\_1 - sub\_metering\_2 - sub\_metering\_3 \quad (1)$$

Since this variable was not in the data, we created it and added it as the target variable.

The variables (columns) already present in the data include:

Columns:

Date: The date (in dd/mm/yyyy format)

Time: The time (in hh:mm:ss format)

Global_active_power: The household's global minute-averaged active power (in kilowatt)

Global_reactive_power: The household's global minute-averaged reactive power (in kilowatt)

Voltage: The minute-averaged voltage (in volt)

Global_intensity: The household's global minute-averaged current intensity (in ampere)

Sub_metering_1: Energy sub-metering No. 1 (in watt-hour of active energy). It corresponds to the kitchen, containing mainly a dishwasher, an oven, and a microwave (hot plates are not electric but gas powered).

Sub_metering_2: Energy sub-metering No. 2 (in watt-hour of active energy). It corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light.

Sub_metering_3: Energy sub-metering No. 3 (in watt-hour of active energy). It corresponds to an electric water-heater and an air-conditioner.

NB: Submetering is the process of measuring the electricity consumption of individual appliances or circuits within a household [4]. In this dataset, we have three submetering columns: `sub_metering_1`, which corresponds to the kitchen, `sub_metering_2`, which corresponds to the laundry room, and `sub_metering_3`, which corresponds to the water-heater and air-conditioner. These columns allow us to examine how different types of appliances contribute to the total energy consumption of the household.

The next thing we did was to explore the data.

2 DATA EXPLORATION

In our exploratory data analysis step, we considered, the missing values present in the data and how significant they were compared to the amount of data present, the sampling rate of the data, the relationships existing between the different variables present in the data (the correlation) and our target variable, the range of values of the variables and the trends, seasonalities found in the data over different sampling periods (i.e. day of the month, month of the year). We then made some assumptions about the data given these observations.

2.1 Statistical Summary of the Variables:

Table 1 below shows the descriptive statistics of the columns or variables in the data.

	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3	total_energy_consumption
count	2.049280e+06	2.049280e+06	2.049280e+06	2.049280e+06	2.049280e+06	2.049280e+06	2.049280e+06	2.049280e+06
mean	1.091615e+00	1.237145e-01	2.408399e+02	4.627759e+00	1.121923e+00	1.298520e+00	6.458447e+00	9.314693e+00
std	1.057294e+00	1.127220e-01	3.239987e+00	4.444396e+00	6.153031e+00	5.822026e+00	8.437154e+00	9.585916e+00
min	7.600000e-02	0.000000e+00	2.232000e+02	2.000000e-01	0.000000e+00	0.000000e+00	0.000000e+00	-2.400000e+00
25%	3.080000e-01	4.800000e-02	2.389900e+02	1.400000e+00	0.000000e+00	0.000000e+00	0.000000e+00	3.800000e+00
50%	6.020000e-01	1.000000e-01	2.410100e+02	2.600000e+00	0.000000e+00	0.000000e+00	1.000000e+00	5.500000e+00
75%	1.528000e+00	1.940000e-01	2.428900e+02	6.400000e+00	0.000000e+00	1.000000e+00	1.700000e+01	1.036667e+01
max	1.112200e+01	1.390000e+00	2.541500e+02	4.840000e+01	8.800000e+01	8.000000e+01	3.100000e+01	1.248333e+02

Table 1: Descriptive statistics of variables

- The household's power consumption varied from 0.076 to 11.122 kilowatts, with an average of 1.092 kilowatts.
- The household's reactive power demand ranged from zero to 1.39 kilowatts, with an average of 0.1237 kilowatts.
- The household's voltage fluctuated from 223.2 to 254.15 volts, with an average of 240.84 volts.
- The kitchen appliances consumed the most power among the sub-metered appliances, with a maximum of 88 watt-hours and an average of 1.1219 watt-hours.
- The laundry room appliances consumed the second most power among the sub-metered appliances, with a maximum of 80 watt-hours and an average of 1.2986 watt-hours.
- The water heater and the air conditioner consumed the least power among the sub-metered appliances, with a maximum of 31 watt-hours and an average of 6.4584 watt-hours.

2.2 Missing values:

We found that the data had missing values, which accounted for nearly 1.25% of the rows. We visualized the missing values using two plots, a heatmap and missing value matrix. From the heatmap in Figure 1 we found that found that the missing points are mostly concentrated in the periods of April 2007, May 2007, and March 2010.

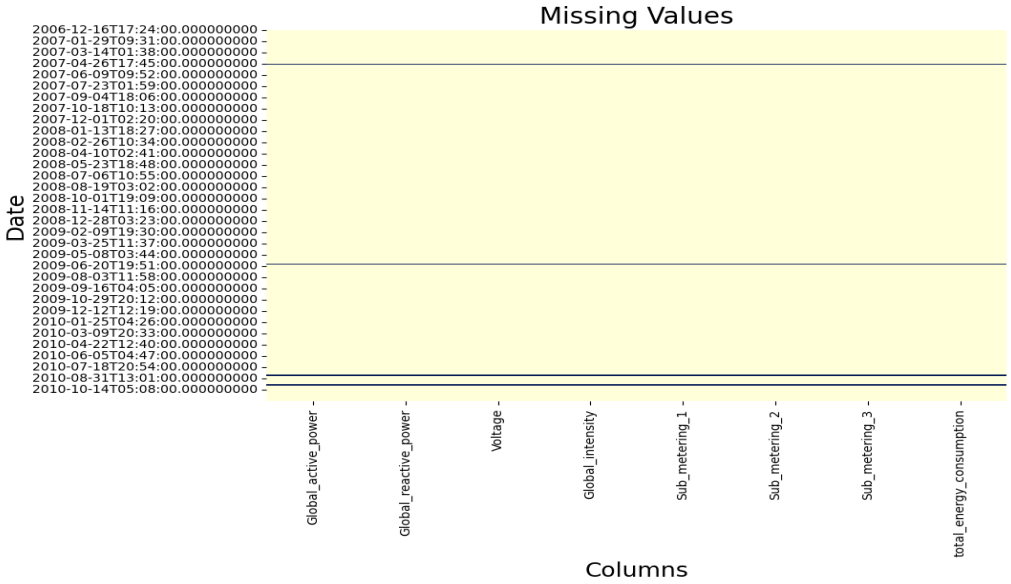


Figure 1: Heatmap showing the missing values and the corresponding dates for the most prominent periods.

From the missing values matrix in figure 2, we observe from the sparkline located at the right-hand side of the figure, however that there a several other less prominent periods of missing measurements within the data.

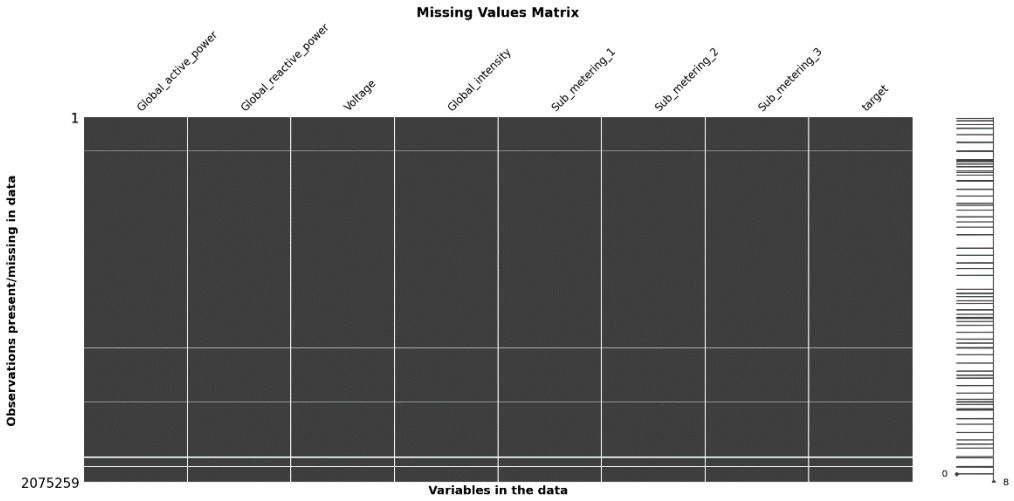


Figure 2: Missing values matrix plot showing the distribution of missing values in each column of the data.

We decided to impute the missing values using the median value of the day where the missing value is found, because we think it is a simple and effective method that preserves the temporal structure of the data.

2.3 Data distribution:

We now observe the distribution of the data in each column, to get more insight into the variability of the columns. We consider the `Global_active_power` column, our target variable, the `total_energy_consumption`, and the 3 submetering columns individually to observe their changes over some aggregate periods of time. Since the data is sampled by the minute, we use **time series boxplots** to visualize how the values for these columns change over the days, days of the week and years and seasons of the year.

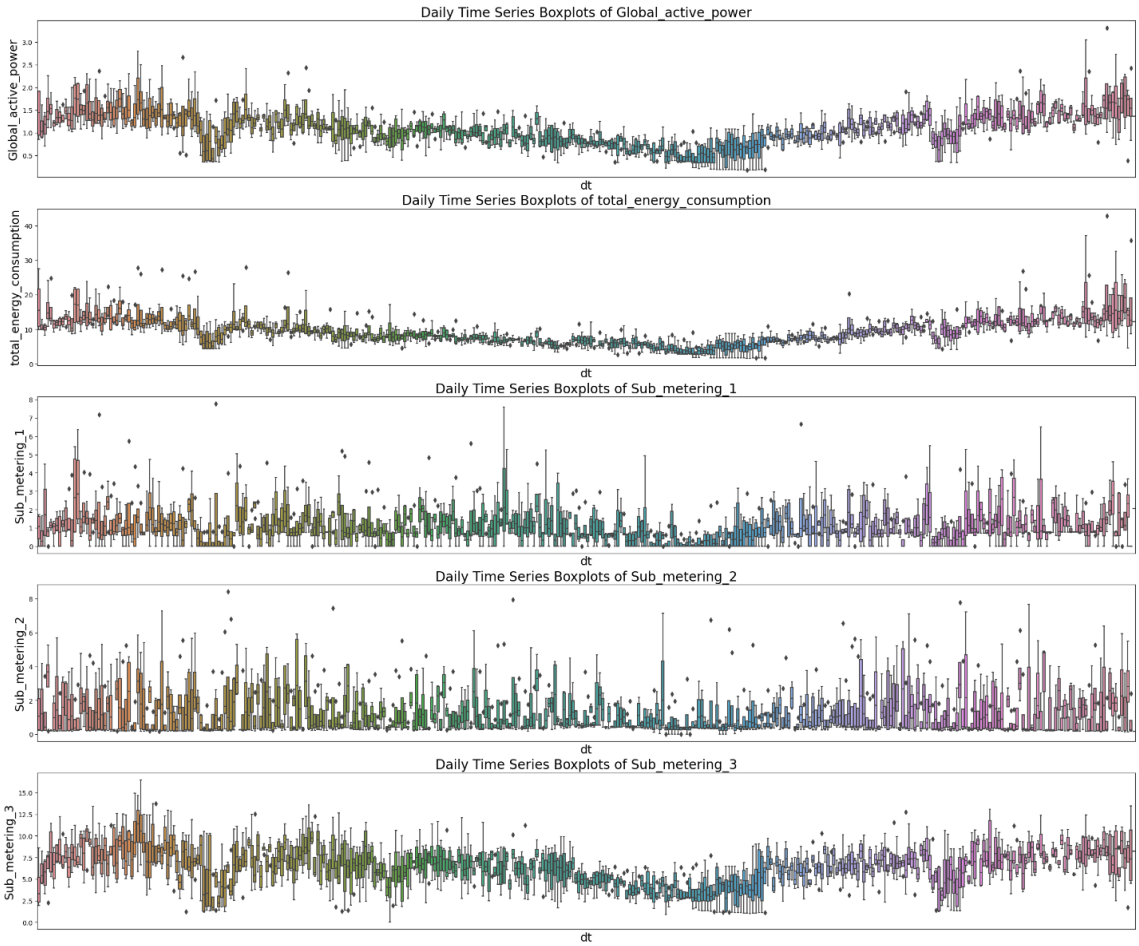


Figure 2: Time series boxplots of electricity consumption by submetering and total energy consumption over days of the week

The subplots reveal to us some outlying high levels of submetering measurements. Specifically, we observe that in the kitchen (`sub_metering_1`) the household typically records a very wide range of electricity consumption in each day and likewise the laundry room (`sub_metering_2`). We also notice that total daily energy consumption has a general motion of falling and then rising over the period during which the data was collected, with relatively smaller ranges in the earlier times and larger ranges in the latter time. We also observe that the total daily energy consumption roughly moves along a similar path as the global daily active power measurement of the household.

We also visualized these distributions over the days of the week and found that they are higher on the weekends than on weekdays. This could be due to the different activities and routines of the household members on weekdays and weekends.

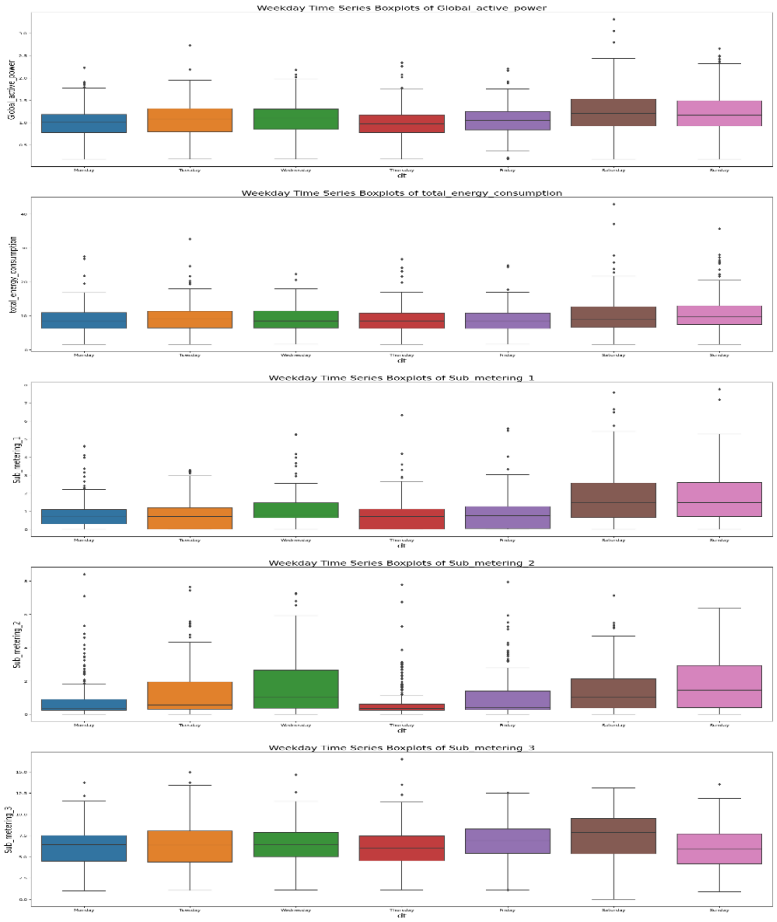
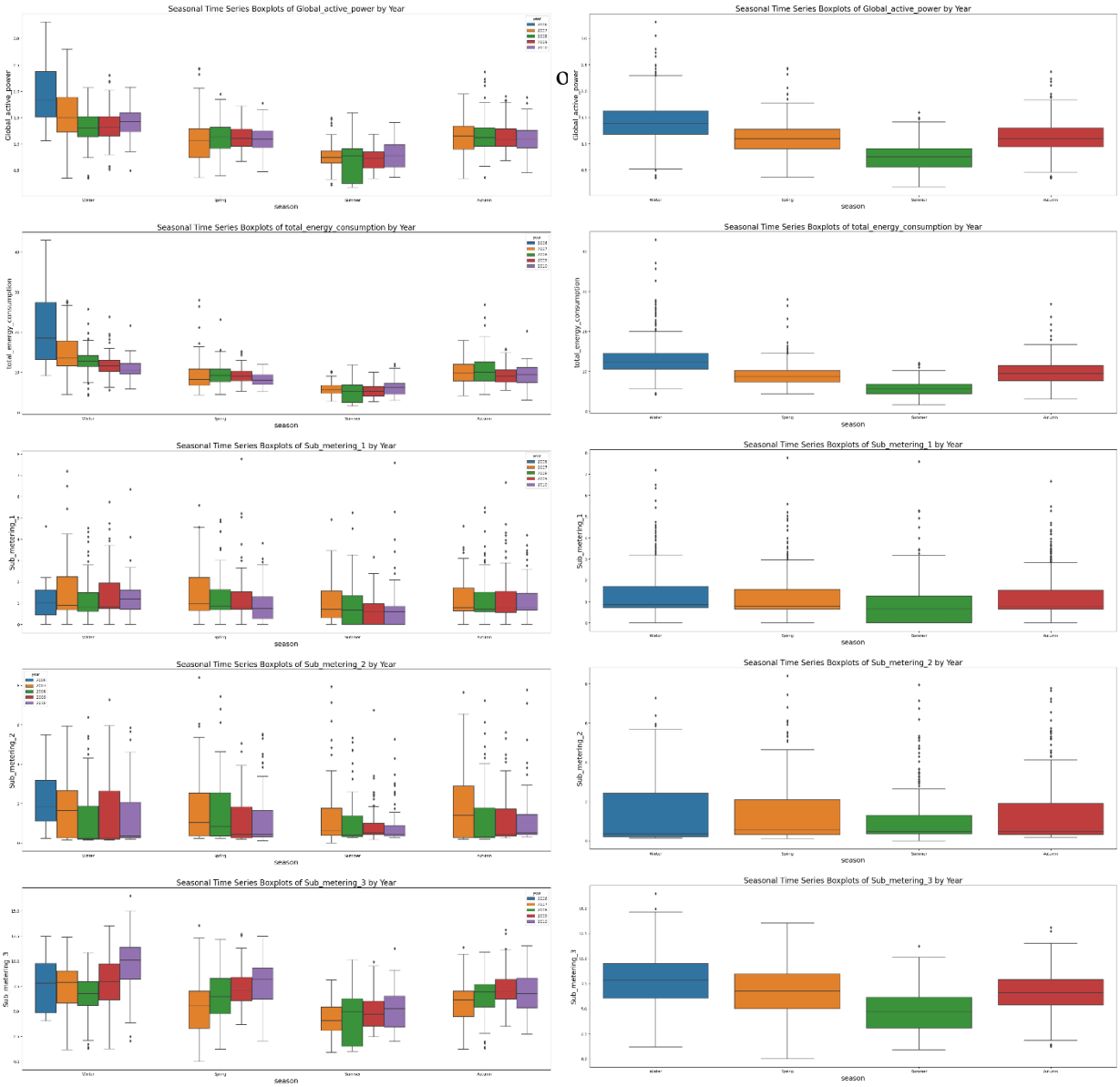


Figure 3: Showing the distribution of energy consumption measurements over the days of the week.

Visualizing the energy consumption distribution over the seasons, we observe that as expected, they are generally higher in the winter than in the summer. This could be due to the increased use of heating systems in the winter and the decreased use of cooling systems in the summer.



These results help us to understand the characteristics and behavior of electric power consumption in the household, and to identify some of the factors that influence it.

Although there were many visualizations, we created to gain a good understanding of the data, we only provided a few of them here.

2.4 Relationships between variables

Using a correlation plot, we study further the variables and how they relate to the total energy consumption variable.

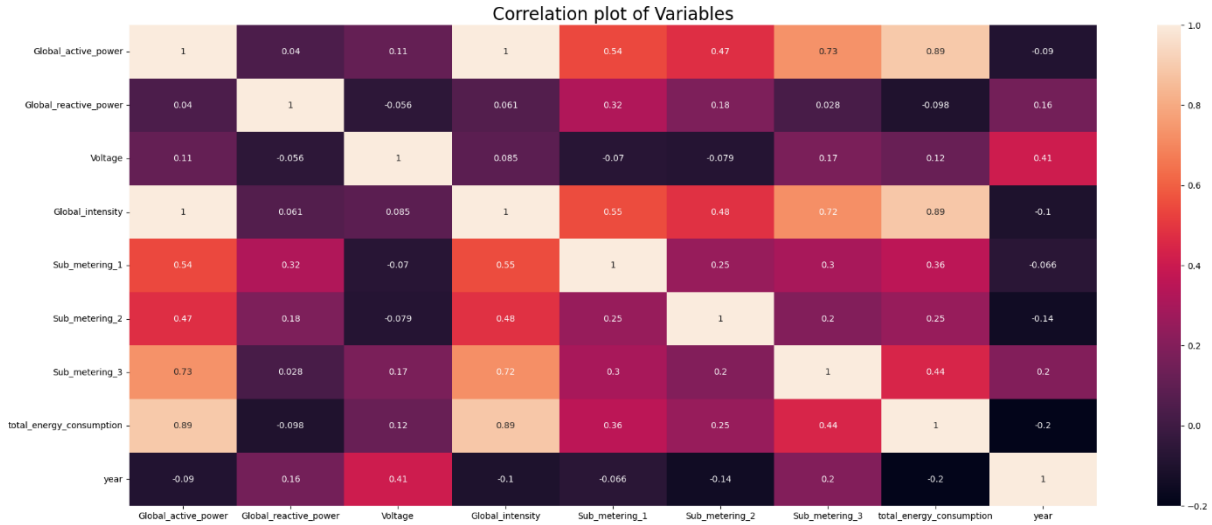


Figure 5: Correlation plots showing the relationships between variables

We observe from the correlation plot that of all the variables provided in the data, the ones with the highest correlation to our variable of interest, which is total energy consumption in decreasing level of importance are global active power, global intensity, sub_metering 3, sub_metering 2 and sub_metering 1.

This is to be expected since we calculated our target variable (total_energy_consumption) using most of these variables, but it also is evidence that a closer look at the trends and seasonality of these quantities could give us some valuable information about the total energy consumption of the household.

2.5 Decomposing Variables into Trends, Seasonality and Residuals

Our next step in understanding the data was to observe the trends, and seasonalities which are present in the columns of the data. Before visualizing these decompositions, however, we first visualized the data for some variables, using the different sampling rates i.e., daily, monthly, quarterly, and yearly.

The figures 6 - 10 below show five plots for each of our 5 selected columns (the top 4 most correlated variables with our target variable in the data and the target variable). The plots show the resampled mean of selected column values over different time periods: day, week, month, quarter, and year. The resampling method is a way of aggregating the data over a specified frequency, such as daily, weekly, monthly, etc. The plots show the trends and patterns of the electric power consumption in the household over time.

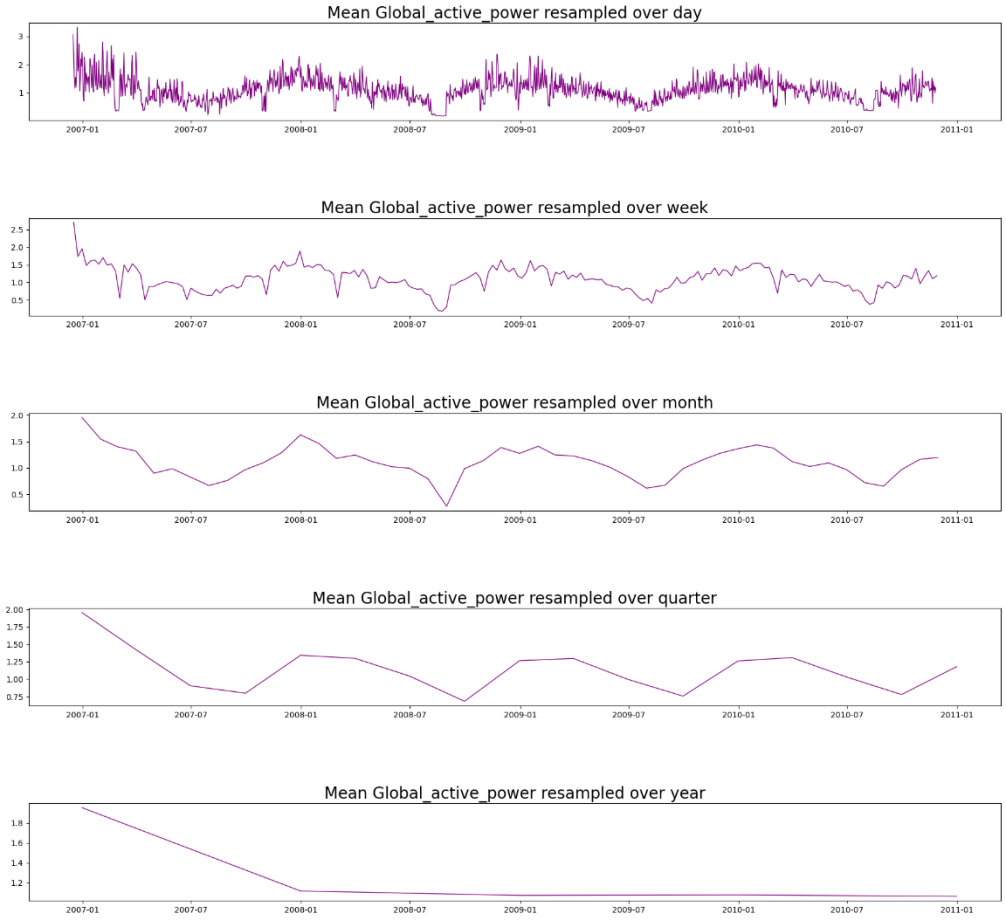


Figure 6: The subplots show the trends and patterns of the global active power over the day, week, month, quarter, and year by the mean.

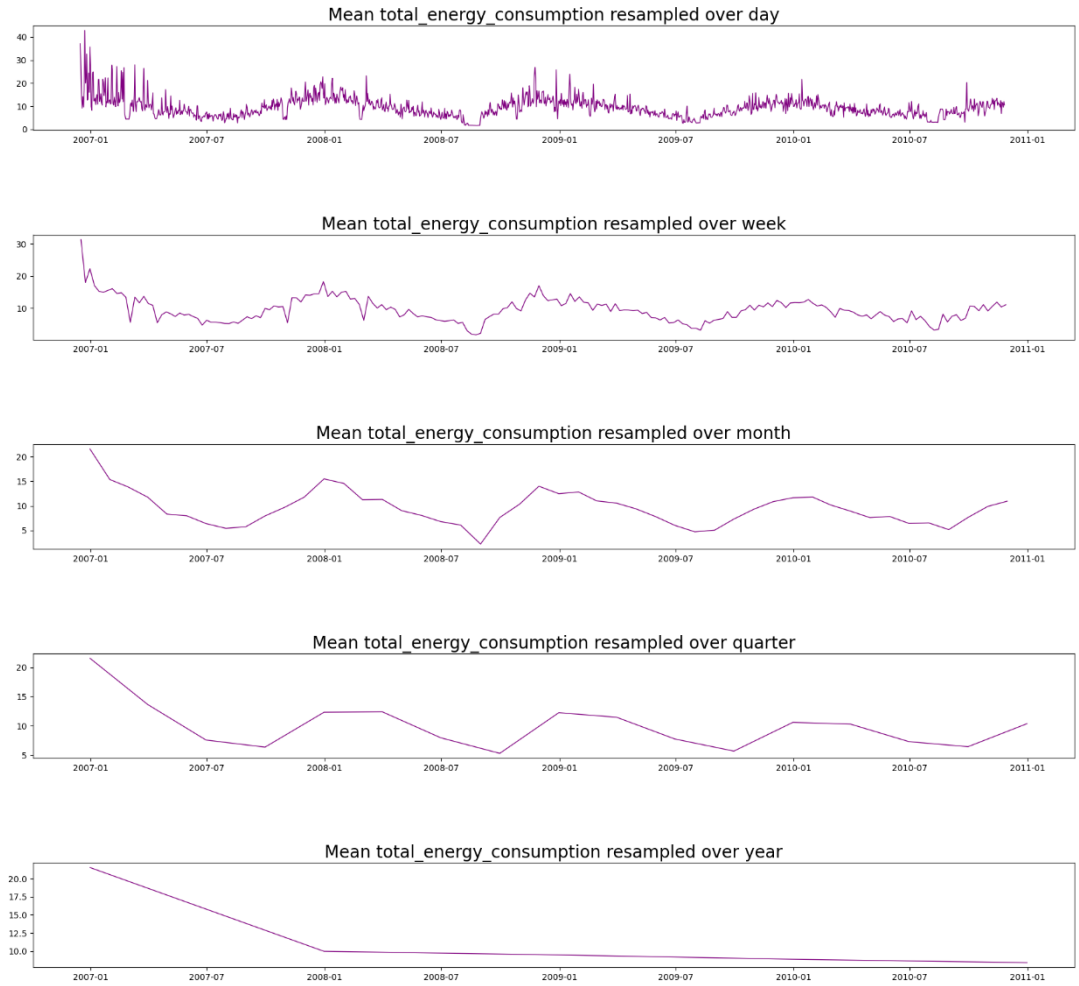


Figure 7: The subplots show the trends and patterns of the total_energy_consumption over the day, week, month, quarter, and year by the mean.

Figure 8: The subplots show the trends and patterns of submetering 1 (kitchen appliances) over the day, week, month, quarter, and year by the mean.

Figure 9: The subplots show the trends and patterns of submetering 2 (washing room appliances) over the day, week, month, quarter, and year by the mean.

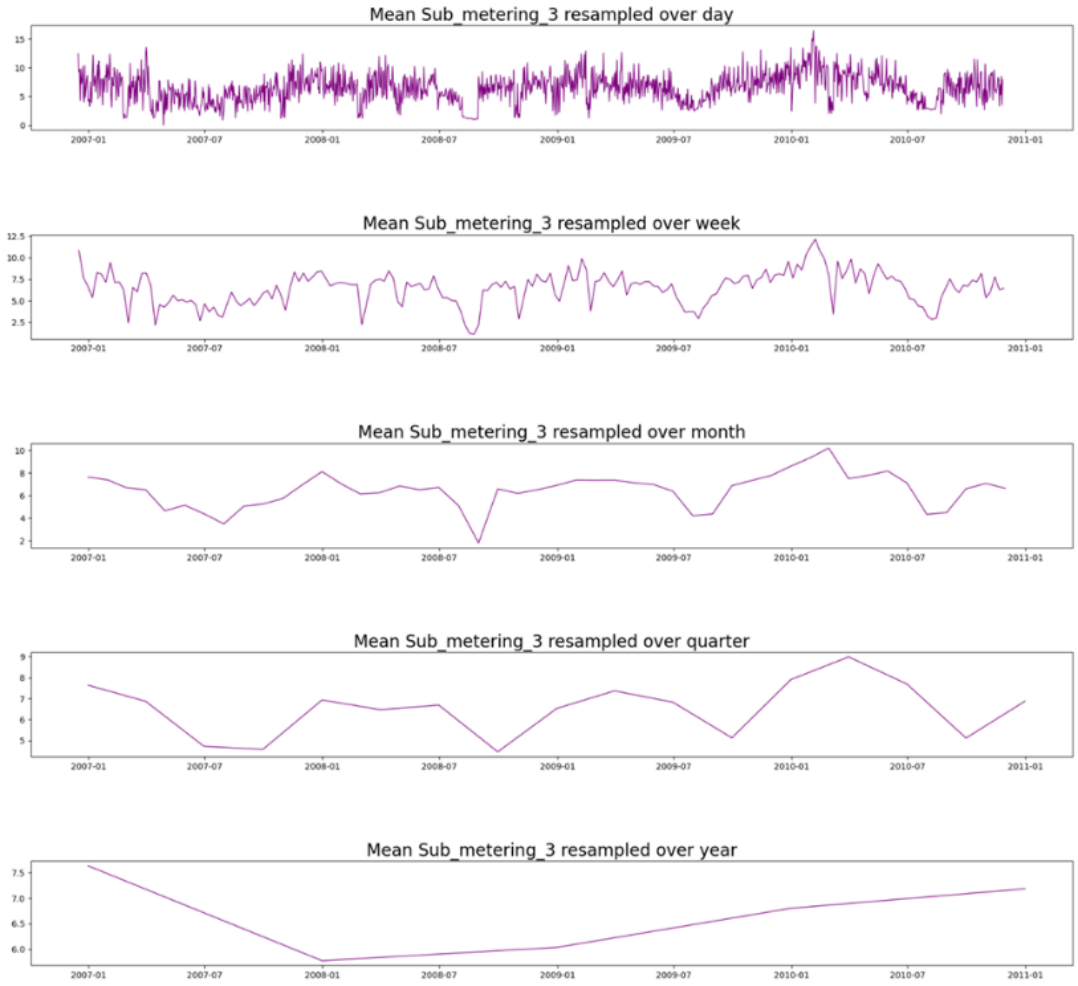


Figure 10: The subplots show the trends and patterns of the sub metering 3 over the day, week, month, quarter, and year by the mean.

From these plots, we see that the energy consumption has a general falling trend of over the yearly periods. We also observe a steady seasonal nature over the quuarters of the year, the months and the weeks.A closer look at these graphs shows that they go according to the seasons (winter, summer, spring, autum and summer).

2.6 Trend Decompositions

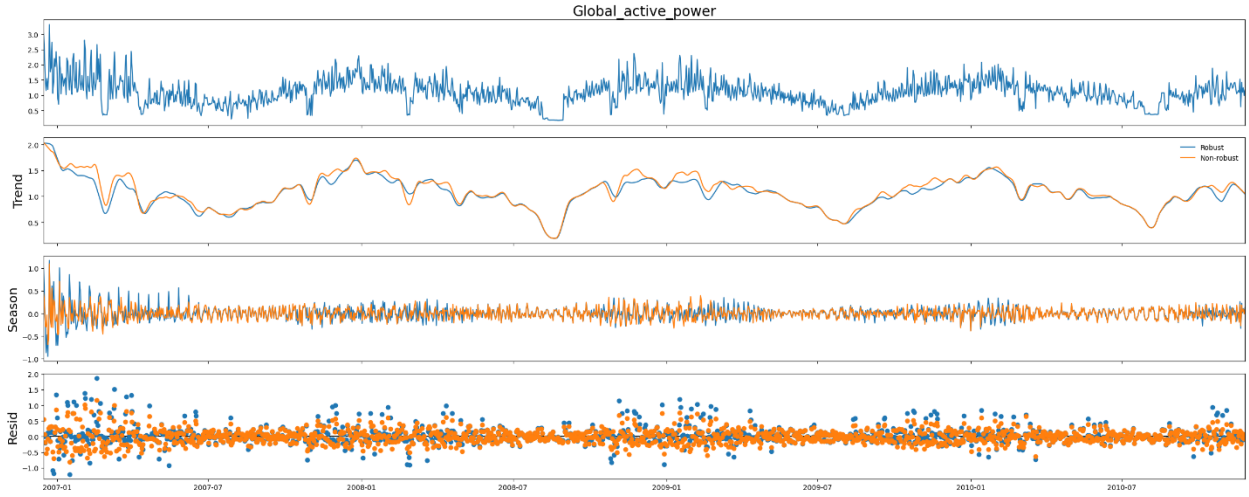


Figure 11: STL decomposition of global active power variable

Figure 12: STL decomposition of total energy consimption variable

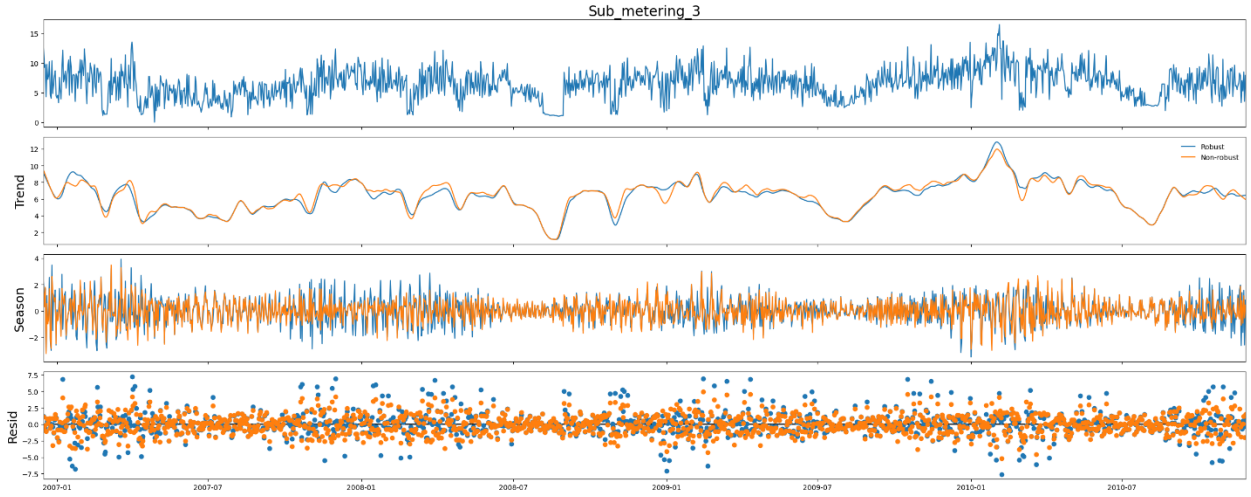


Figure 13: STL decomposition of submetering 3 variable

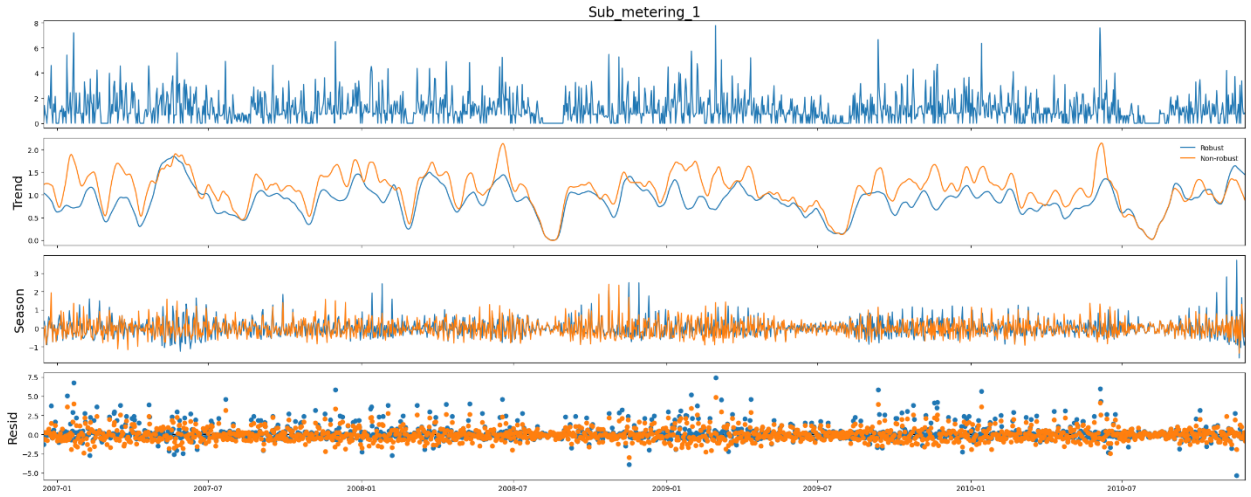


Figure 14: STL decomposition of submetering 1 variable

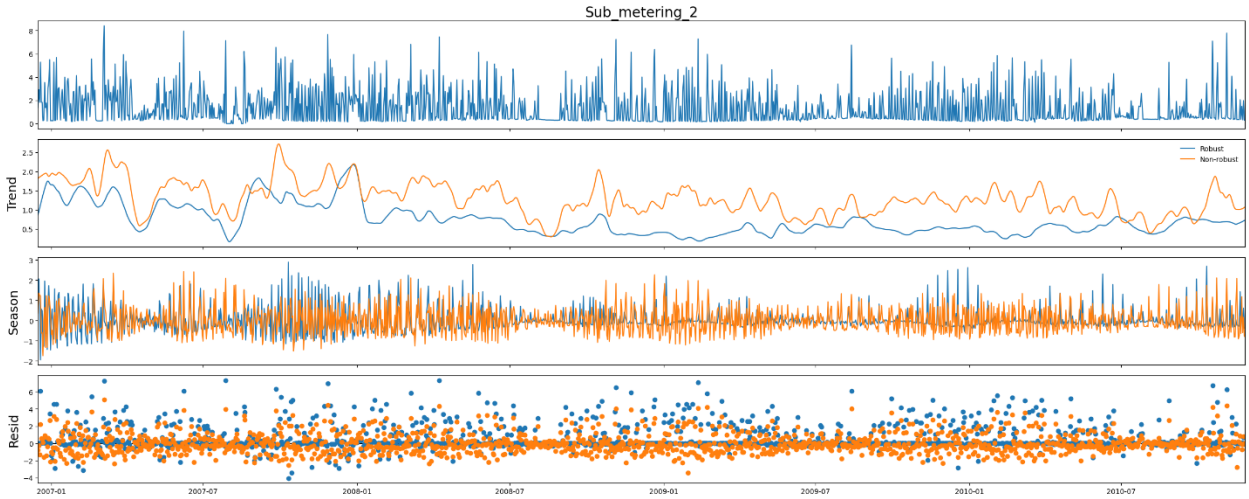


Figure 15: STL decomposition of submetering 2 variable

We observe that these trend decomposition plots bare an interesting resemblance to the previous sets of plots we made in section 2.5. We also observe that the residual plots for each of the variables have some slight level of structure, implying that there may be some other factors affecting the variables which have not been accounted for by the seasonalities we have plotted so far.

We continue to explore and draw more insights on these obsevrations.