CAPSTONE PROJECT

Machine Learning on Time-Series Analysis

Prediction on Household Power Consumption Usage using ARIMA & SARIMA Prediction Model

Najibah binti Ali

Data Science Analytical

Cohort 3 Session 2024

TABLE OF CONTENTS

1. EXECUTIVE SUMMARY

The objective of this capstone project is to develop a predictive model for future household power consumption using time series analysis techniques, specifically the ARIMA (*Auto-regressive Integrated Moving Average*) & SARIMA (*Seasonal Auto-Regressive Integrated Moving Average*) model. Accurate predictions can significantly enhance energy management and planning, leading to more efficient resource allocation and cost savings. The dataset, sourced from Kaggle, comprises detailed records of household power consumption collected over several months in 2007. This project adheres to the SMART framework: it is *Specific* in predicting household power usage using the ARIMA model, *Measurable* through accuracy and error metrics, *Achievable* with clear guidance and organized data, *Relevant* for utility companies to optimize power generation and distribution as well as for the public to minimize over-consumption, and *Time-bound* with a completion deadline before July 1st, 2024.

To evaluate the model, Root Mean Squared Error (*RMSE*) and Mean Absolute Error (*MAE*) were selected as performance metrics. Following extensive data cleaning, munging, formatting, and exploration, the model was fine-tuned using Auto-ARIMA, demonstrating good predictive accuracy on the test data. However, several risks and limitations must be acknowledged. Data quality is a critical risk, as incomplete or incorrect data entries can affect model accuracy. Risks also include multicollinearity and data relevancy, with potential issues if the dataset is unsuitable for the project objective. The model's limitations include its training on a specific dataset, which may restrict its generalizability. Additionally, external factors such as weather conditions, holidays, and socioeconomic changes are not considered, which could impact power consumption patterns. This project is built on several assumptions: the data is cleaned and validated, the data distribution is representative, historical consumption patterns will continue, and there are no significant changes in household behaviour or external factors influencing power consumption. Adhering to these assumptions and mitigating the identified risks will be crucial for the successful completion and application of the predictive model.

In summary, this project aims to create an ARIMA-based and SARIMA-based predictive model for household power consumption using a detailed dataset from Kaggle. The model’s development and performance will be measured using RMSE and MAE. While there are risks and limitations related to data quality and external factors, the project is well-defined, achievable, and relevant, with a clear timeline for completion.

2. SCIENTIFIC REPORT

1.0 Abstract

This study aims to forecast household power consumption using time series analysis. The dataset was processed and analysed to identify trends and patterns. An ARIMA and SARIMA model was employed to predict future consumption, and its performance was evaluated using various error metrics. The results demonstrate the model's capability to provide accurate short-term forecasts, with implications for improved energy management. These predictions can aid in efficient energy management and planning.

2.0 Introduction

In addition, this project serves as a practical application of the concepts have been studying in data science class. It provides exposure to the complete process of data preprocessing, exploratory data analysis (EDA), and the implementation of a machine learning model. This hands-on experience is intended to enhance the skills in handling real-world datasets and producing technical reports for stakeholders.

The prediction of household power consumption is a vital task for energy providers and policymakers. Accurate forecasts enable better demand planning, load balancing, and resource allocation, contributing to cost savings and efficiency. This project utilizes time series analysis to forecast future power usage based on historical data; by implementing and optimizing ARIMA & SARIMA model, the study seeks to provide reliable consumption forecasts, thereby supporting improved energy management strategies.

3.0 Materials and Methods

I. Materials

- **Dataset**: Household power consumption data from Kaggle.
- **Software**: Jupyter Notebook, Python libraries (pandas, numpy, matplotlib, seaborn, statsmodels, sklearn, pmdarima).

II. Methods

a) Data Acquisition

The dataset was downloaded from Kaggle website in CSV format and loaded into a Jupyter Notebook for analysis.

b) Data Cleaning and Munging

- **Handling Missing Values**: Missing values were identified and imputed filling method either with mean or sum.
- **Data Resampling**: The data was resampled to weekly intervals to reduce noise and capture consumption patterns.
- **DateTime Format**: The 'Date' and 'Time' column are converted to a DateTime format to facilitate time series operations and ensure accurate time indexing.
- **Data Type Conversion**: Numerical columns that were initially read as objects were converted to floats to allow for mathematical operations and analysis.

c) Exploratory Data Analysis (EDA)

- **Non-Graphical Analysis**: Groupby operations and summary statistics were used to understand the data distribution and trends.
- **Graphical Analysis**: Various plots, including time series plots, histograms, box plots, and correlation heatmaps, were generated to visualize data patterns and relationships.

d) Stationarity Check

The Augmented Dickey-Fuller (ADF) test was applied to check for stationarity. The data was transformed using differencing until stationarity was achieved.

e) Model Selection and Implementation

- **ARIMA & SARIMA Model**: The model was selected for its effectiveness in handling time series data with trends and seasonality.
- **Auto-ARIMA**: The Auto-ARIMA function from the pmdarima library was used to automatically determine the best parameters (p, d, q, P, D,Q, S) for the model.

f) Model Evaluation

The model was evaluated using MAE and RMSE to assess its predictive accuracy. The p, d and q parameter are chosen based on the function with low AIC (Akaike Information Criterion) and produce the lowest RMSE.

g) Prediction Reading

After training the ARIMA and SARIMA model on the cleaned and pre-processed dataset, predictions were generated beyond the current available data. This involved:

- **Model Training**: The ARIMA & SARIMA model was fitted to the train set of historical data, capturing the underlying trends and seasonal patterns.
- **Forecast Generation**: Using the fitted model, forecasts were generated on test set for months beyond the timeline. These predictions were evaluated for error.
- **Result Visualization**: The forecasted values were plotted alongside the historical data to visualize the predicted consumption trends and assess the model's performance.

4.0 Results

Through this project, I embarked on a comprehensive journey to apply data science and machine learning techniques to a real-world problem, starting from scratch.

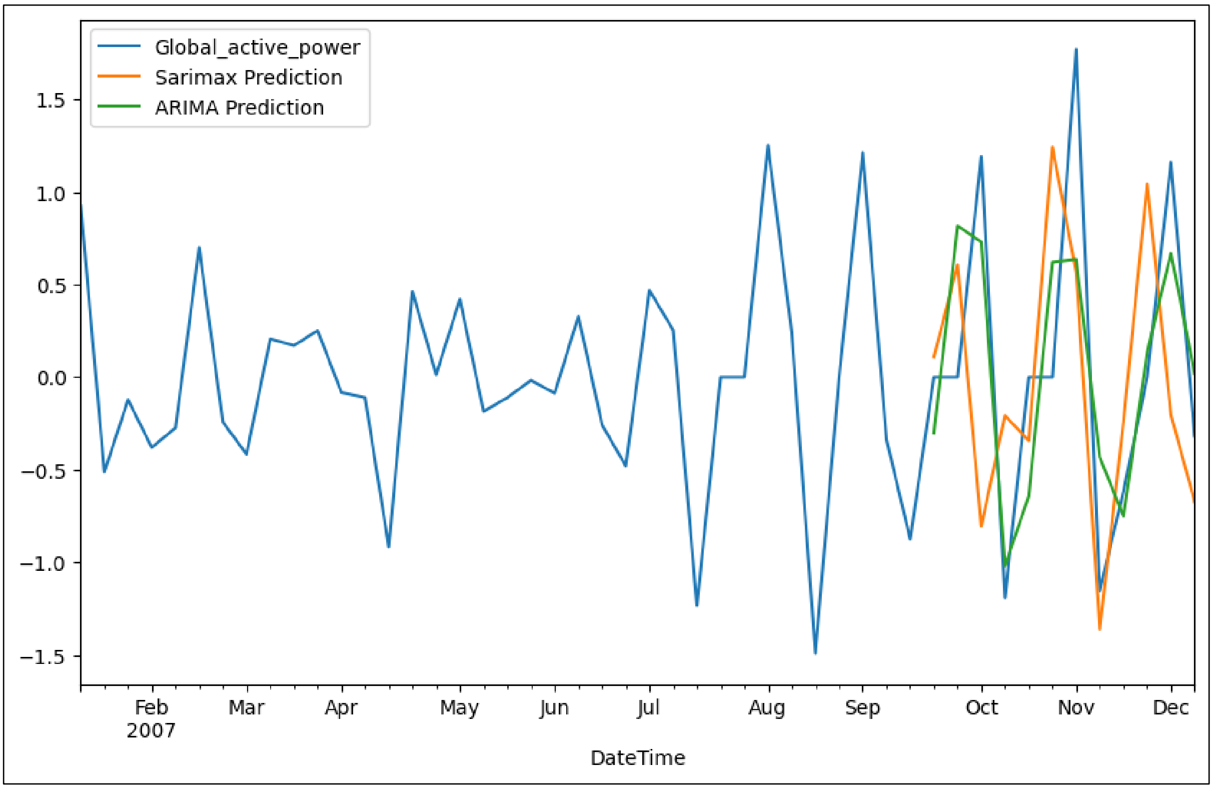


Figure 1: The comparison between ARIMA & SARIMA prediction modelling on the original Global active power plot.

Table 1: The Global active power prediction for ARIMA modelling

Date	ARIMA Predictions for Global active power(kW)
2007-07-01	1.518
2007-08-01	1.418
2007-09-01	2.338
2007-10-01	2.518
2007-11-01	1.618
2007-12-01	1.318
2008-01-01	2.098
2008-02-01	2.498
2008-03-01	1.778
2008-04-01	1.318
2008-05-01	1.908
2008-06-01	2.438
2008-07-01	1.908
2008-08-01	1.358
2008-09-01	1.758
2008-10-01	2.348
2008-11-01	2.018
2008-12-01	1.428
2009-01-01	1.638
2009-02-01	2.238
2009-03-01	2.078
2009-04-01	1.498
2009-05-01	1.548
2009-06-01	2.118
2009-07-01	2.118
2009-08-01	1.588
2009-09-01	1.488
2009-10-01	1.988
2009-11-01	2.118
2009-12-01	1.668
2010-01-01	1.458
2010-02-01	1.868
2010-03-01	2.088
2010-04-01	1.738
2010-05-01	1.458
2010-06-01	1.768
2009-01-01	1.638
2009-02-01	2.238
2009-03-01	2.078

Table 2: The Global active power prediction for SARIMA modelling

Date	SARIMA Predictions for Global active power(kW)
2007-07-01	3.283
2007-07-08	2.621
2007-07-15	2.302
2007-07-22	2.321
2007-07-29	3.209
2007-08-05	4.144
2007-08-12	4.055
2007-08-19	2.750
2007-08-26	2.595
2007-09-02	3.350
2007-09-09	3.394
2007-09-16	2.751
2007-09-23	2.860
2007-09-30	3.466
2007-10-07	2.661
2007-10-14	2.453
2007-10-21	2.110
2007-10-28	3.353

5.0 Discussion

The results highlight the ARIMA and SARIMA model in forecasting household power consumption. However, the model's performance may be impacted by external factors not included in the dataset. Future work could involve integrating additional variables, such as weather conditions and socioeconomic factors, to enhance predictive accuracy. Moreover, exploring other advanced time series models like LSTM could provide further improvements. Here are the key steps and achievements:

- **Data Acquisition**: I learned how to source data from Kaggle, a reliable platform that provides various datasets for analysis.
- **Data Preprocessing**: I gained hands-on experience in data cleaning, munging and formatting. This included handling missing value, resampling the data to weekly intervals to reduce noise, detecting and mitigating outliers using statistical methods and visual inspections, converting the 'Date' and ‘Time’ columns to DateTime format, and changing numerical columns from object type to float to facilitate modelling operations.
- **Exploratory Data Analysis (EDA)**: I performed both non-graphical and graphical EDA to understand the underlying patterns in the data. This involved using groupby methods, creating correlation heatmaps, and generating various plots to visualize trends in the dataset.
- **Data Preparation for Modelling**: I ensured that the dataset was suitable for time series modelling by selecting useful data, checking for stationarity and applying differencing techniques where necessary. This step was crucial for the accurate application of the ARIMA & SARIMA model.
- **ARIMA & SARIMA Modelling**: I successfully implemented the prediction model to analyse the time series data. Using Auto-ARIMA, I identified the best parameters for the model, ensuring optimal performance by low AIC value. The model was trained on training set and used to generate forecasts on the testing set, achieving the project's objective.
- **Prediction and Evaluation**: The ARIMA & SARIMA model provided low error predictions for future household power consumption. The results were evaluated

using MAE and RMSE metrics, demonstrating the model's reliability in capturing trends and seasonal patterns.

The ARIMA and SARIMA model, optimized using Auto-ARIMA, provided accurate predictions of household power consumption. The evaluation metrics indicated good performance, with the model capturing both the trend and seasonal components of the data. The MAE and RMSE values were within acceptable ranges, confirming the model's reliability.

The entire process and results were documented and presented in Jupyter Notebook. To share my work with the wider community, I have uploaded the project repository to GitHub. Additionally, I write a Medium article detailing my first machine learning project, sharing insights and learnings with other beginners in the field of data science.

- **GitHub Repository**: [Link to GitHub Repository]
- **Medium Article**: Predicting Household Power Consumption Using ARIMA and SARIMA Models: A CRISP-DM Approach

6.0 Acknowledgements

I would like to express my deepest gratitude to my instructor, whose invaluable guidance, support, and encouragement have been instrumental throughout this project. Their expertise and patience have significantly contributed to my understanding and application of data science principles. I would also like to thank the Peoplelogy team, that provided the materials and classes that have been foundational to my journey as a data science beginner. The comprehensive resources and structured learning environment have been crucial in developing my skills and confidence in this field. Your contributions have made this project possible and have prepared me for future endeavours in the realm of data science.

7.0 References

Hebbar, N. (2020, Sep 18). “Time Series Forecasting With ARIMA Model in Python for Temperature Prediction”. Medium. https://medium.com/swlh/temperature-forecasting-with-arima-model-in-python-427b2d3bcb53

Howell, E. (2023, Jan 31). “What is ARIMA? An introduction to the ARIMA forecasting model and how to use it for time series.”. Medium. https://towardsdatascience.com/how-to-forecast-with-arima-96b3d4db111a

Kumar, R. (2023, Apr 1). “Let’s talk🦜about your first Time series ARIMA model 💹: Part 1”. Medium. https://blog.devops.dev/lets-talk-about-your-first-arima-model-cbfdcba1749e

L. & J. (2022, Aug 25). “How to build ARIMA models in Python for time series prediction”. JustintoData. https://www.justintodata.com/arima-models-in-python-time-series-prediction/

Monigatti, L. (2022, Aug 2). “Interpreting ACF and PACF Plots for Time Series Forecasting”. Medium. https://towardsdatascience.com/interpreting-acf-and-pacf-plots-for-time-series-forecasting-af0d6db4061c

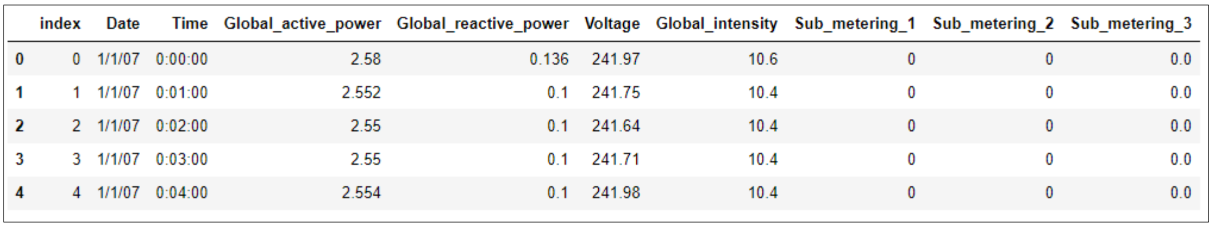
Umrajkar, V. (2022). “Electricity Consumption-Time Series Analysis”. Kaggle. https://www.kaggle.com/code/vedumrajkar/electricity-consumption-time-series-analysis/notebook

Wainaina, P. (2023, Oct 24). “The Complete Guide to Time Series Forecasting Models”. Medium. https://medium.com/@wainaina.pierre/the-complete-guide-to-time-series-forecasting-models-ef9c8cd40037

8.0 Appendix

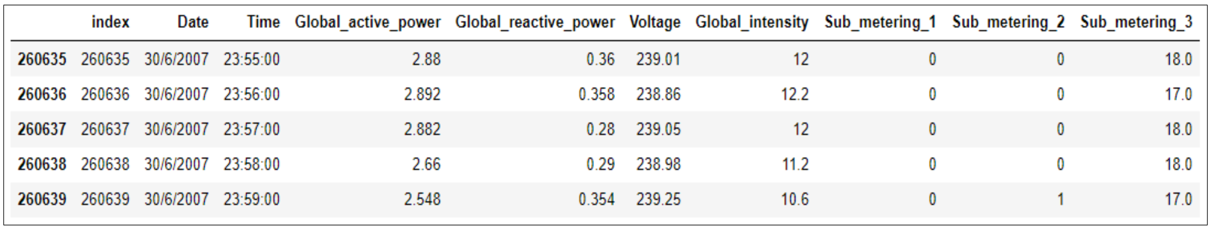
a. Understanding & Preparation of Data

The first 5rows of raw data



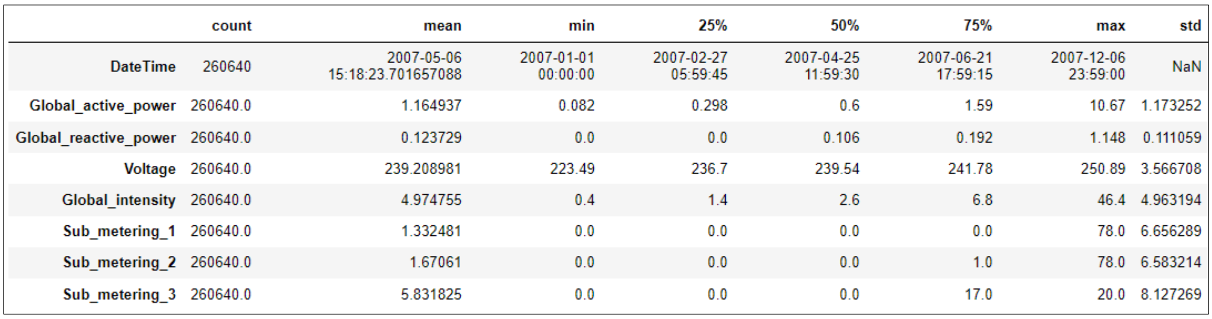
	index	Date	Time	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3
0	0	1/1/07	0:00:00	2.58	0.136	241.97	10.6	0	0	0.0
1	1	1/1/07	0:01:00	2.552	0.1	241.75	10.4	0	0	0.0
2	2	1/1/07	0:02:00	2.55	0.1	241.64	10.4	0	0	0.0
3	3	1/1/07	0:03:00	2.55	0.1	241.71	10.4	0	0	0.0
4	4	1/1/07	0:04:00	2.554	0.1	241.98	10.4	0	0	0.0

The last 5 rows of raw data: All the features are cleaned and formatted to numeric.



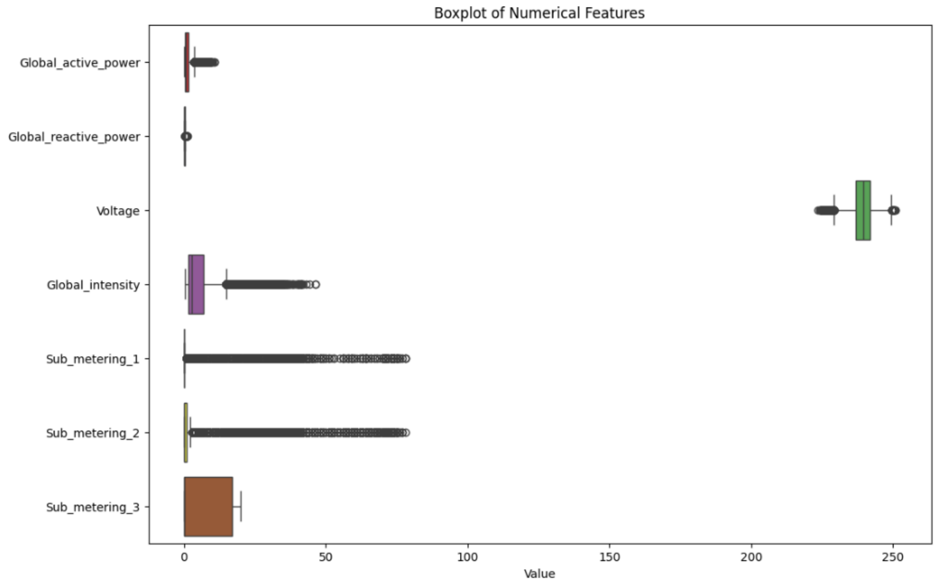
	index	Date	Time	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3
260635	260635	30/6/2007	23:55:00	2.88	0.36	239.01	12	0	0	18.0
260636	260636	30/6/2007	23:56:00	2.892	0.358	238.86	12.2	0	0	17.0
260637	260637	30/6/2007	23:57:00	2.882	0.28	239.05	12	0	0	18.0
260638	260638	30/6/2007	23:58:00	2.66	0.29	238.98	11.2	0	0	18.0
260639	260639	30/6/2007	23:59:00	2.548	0.354	239.25	10.6	0	1	17.0

Description for the features

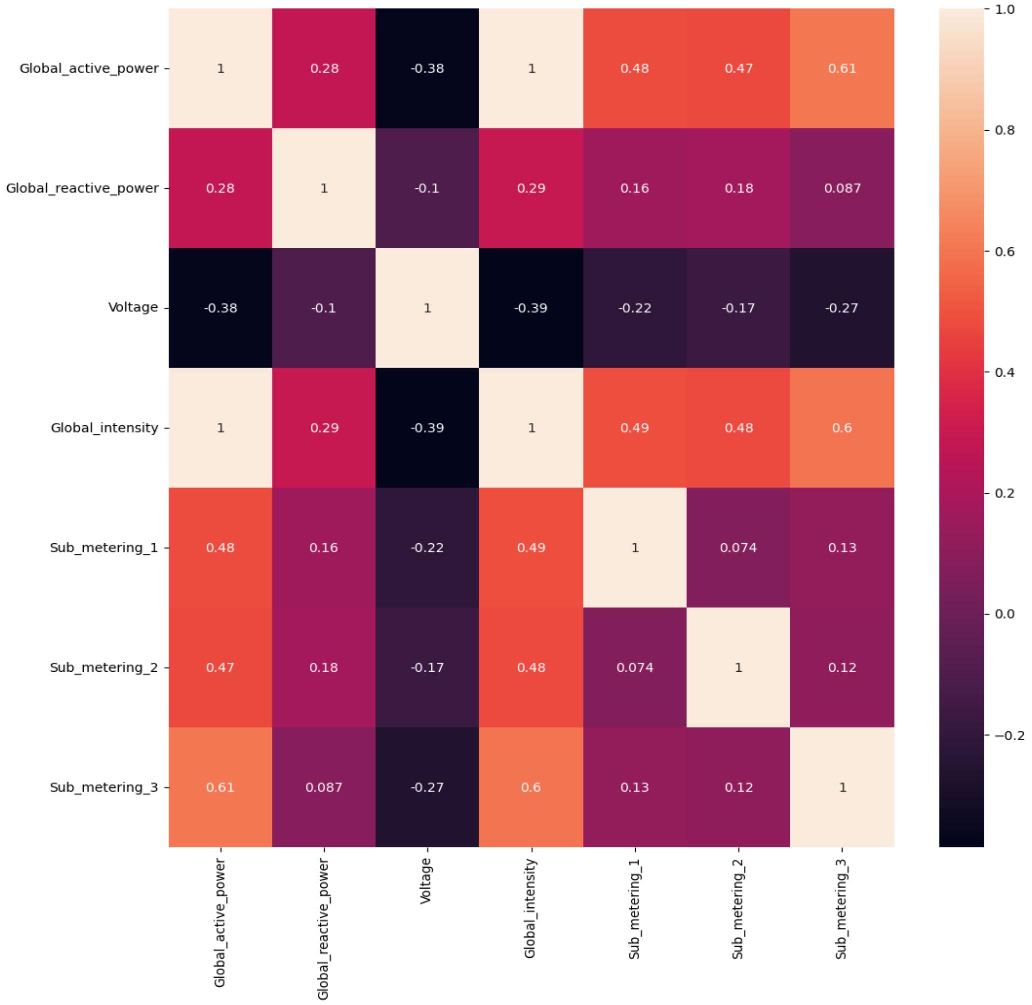


	count	mean	min	25%	50%	75%	max	std
DateTime	260640	2007-05-06 15:18:23.701657088	2007-01-01 00:00:00	2007-02-27 05:59:45	2007-04-25 11:59:30	2007-06-21 17:59:15	2007-12-06 23:59:00	NaN
Global_active_power	260640.0	1.164937	0.082	0.298	0.6	1.59	10.67	1.173252
Global_reactive_power	260640.0	0.123729	0.0	0.0	0.106	0.192	1.148	0.111059
Voltage	260640.0	239.208981	223.49	236.7	239.54	241.78	250.89	3.566708
Global_intensity	260640.0	4.974755	0.4	1.4	2.6	6.8	46.4	4.963194
Sub_metering_1	260640.0	1.332481	0.0	0.0	0.0	0.0	78.0	6.656289
Sub_metering_2	260640.0	1.67061	0.0	0.0	0.0	1.0	78.0	6.583214
Sub_metering_3	260640.0	5.831825	0.0	0.0	0.0	17.0	20.0	8.127269

Boxplot for outliers

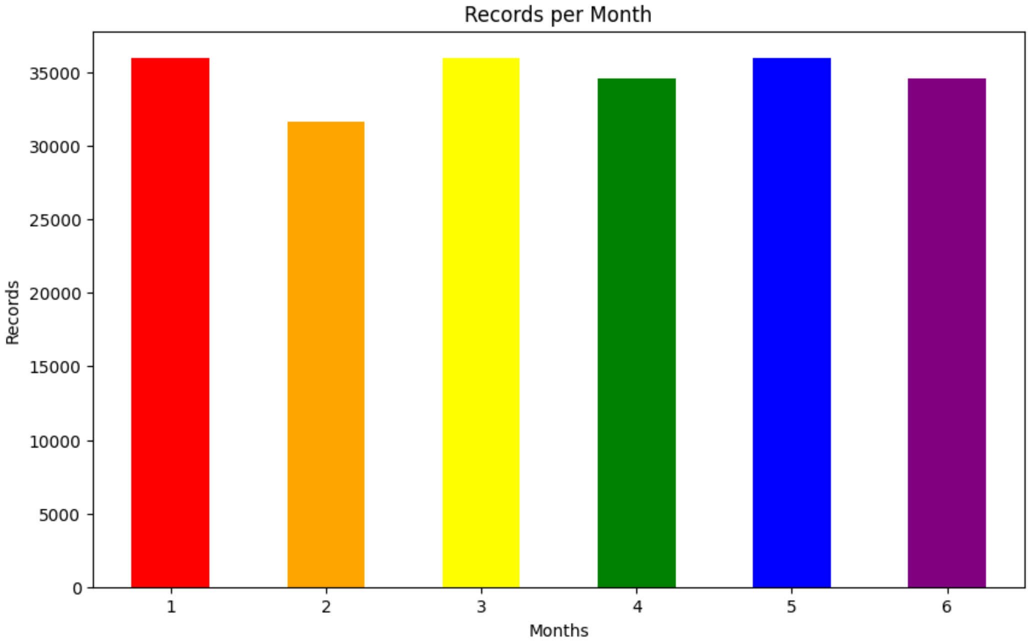


Heatmap of features correlation

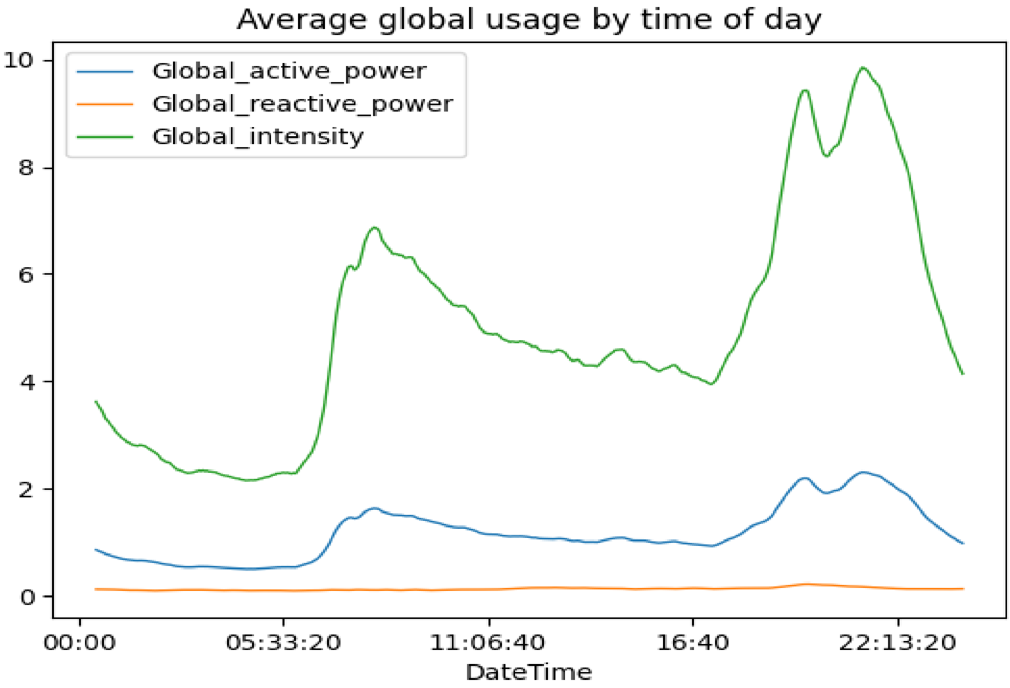


	Global_active_power	Global_reactive_power	Voltage	Global_intensity	Sub_metering_1	Sub_metering_2	Sub_metering_3
Global_active_power	1	0.28	-0.38	1	0.48	0.47	0.61
Global_reactive_power	0.28	1	-0.1	0.29	0.16	0.18	0.087
Voltage	-0.38	-0.1	1	-0.39	-0.22	-0.17	-0.27
Global_intensity	1	0.29	-0.39	1	0.49	0.48	0.6
Sub_metering_1	0.48	0.16	-0.22	0.49	1	0.074	0.13
Sub_metering_2	0.47	0.18	-0.17	0.48	0.074	1	0.12
Sub_metering_3	0.61	0.087	-0.27	0.6	0.13	0.12	1

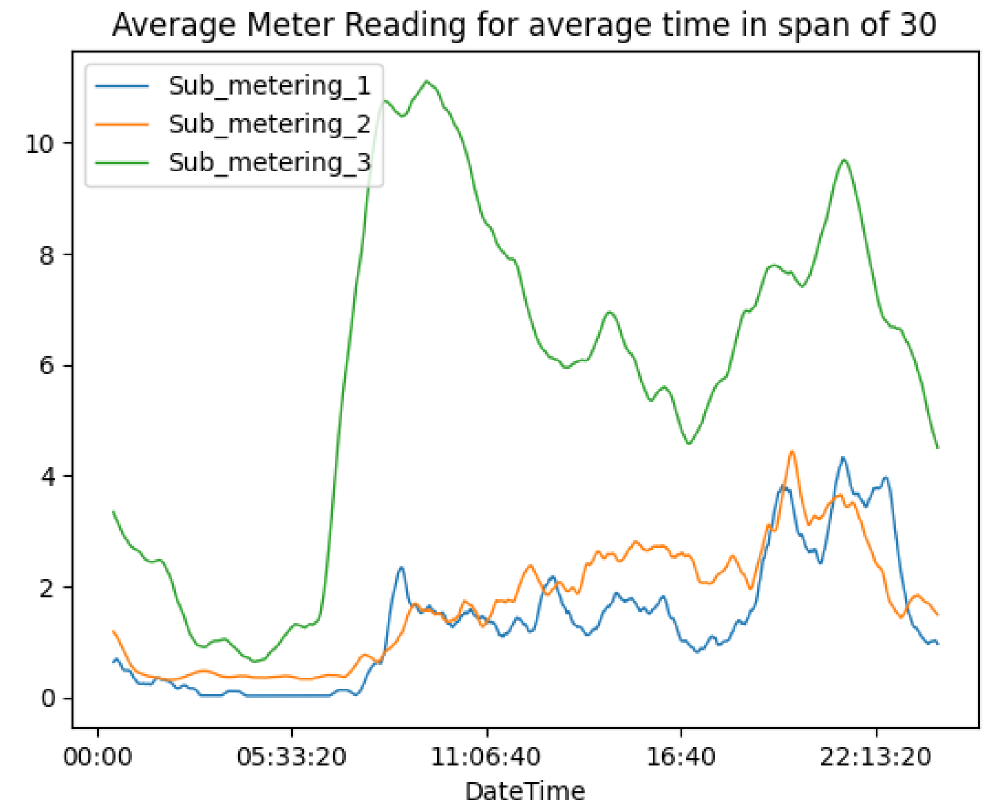
Plot for records taken for each month.



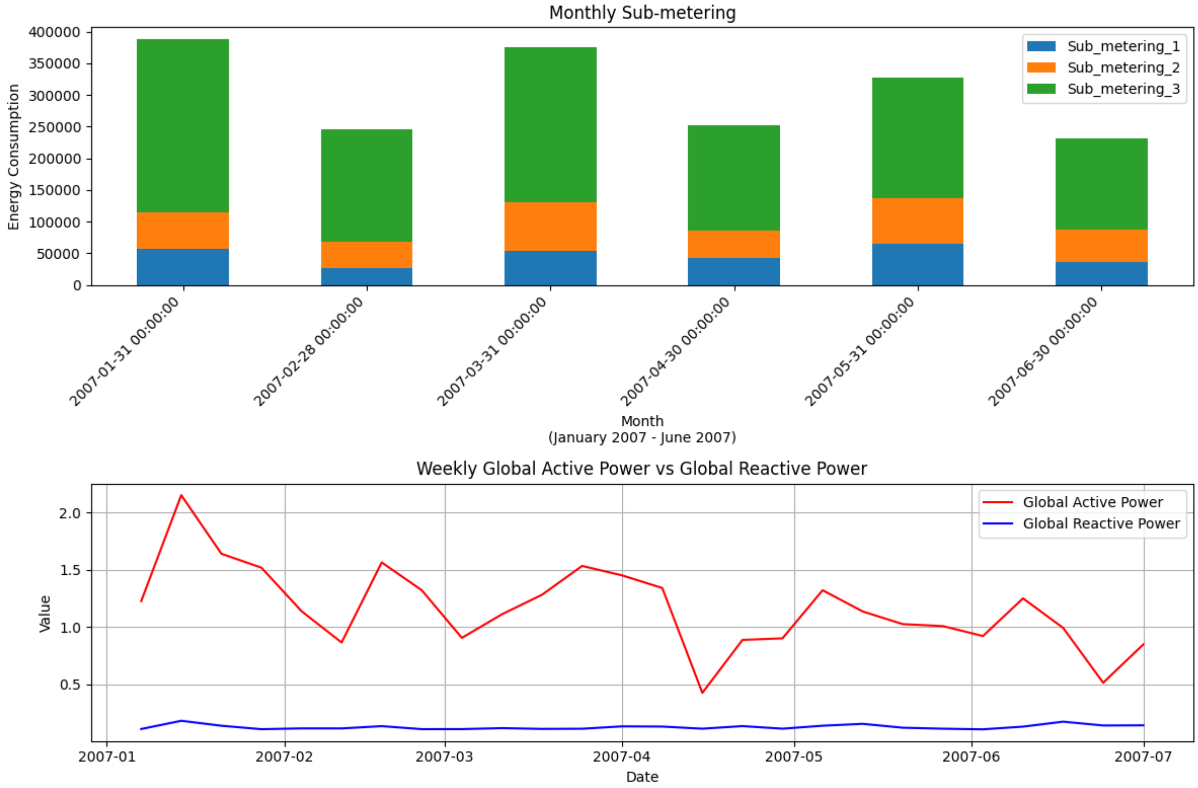
Average global usage by time of the day, rolling 30 readings



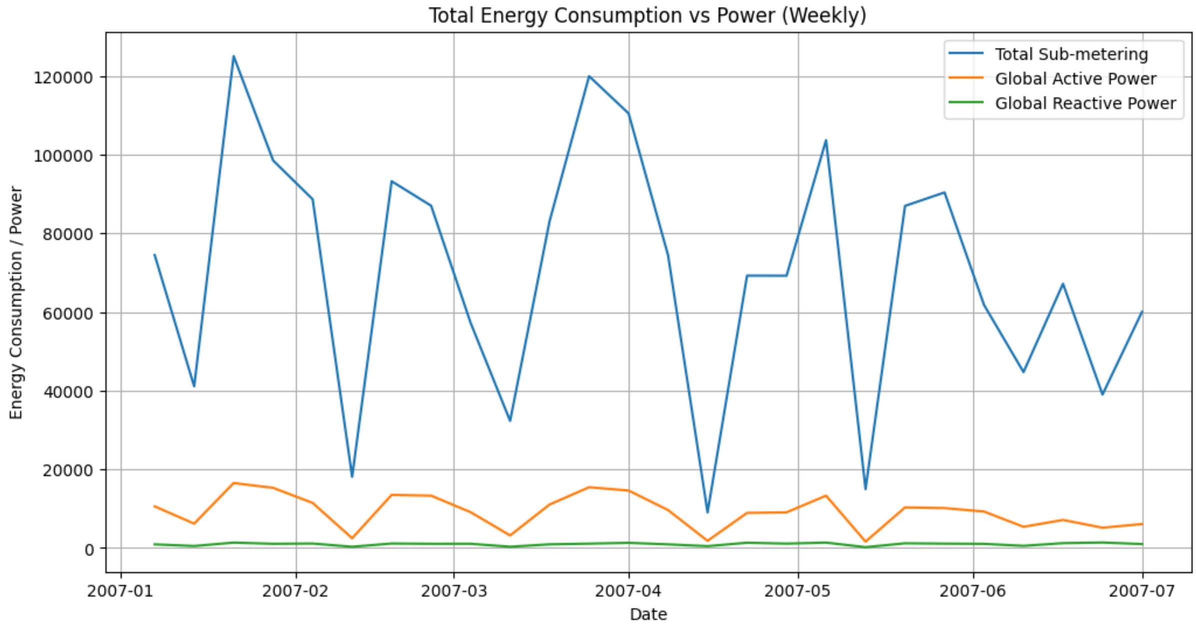
Average meter reading in span of 30 readings



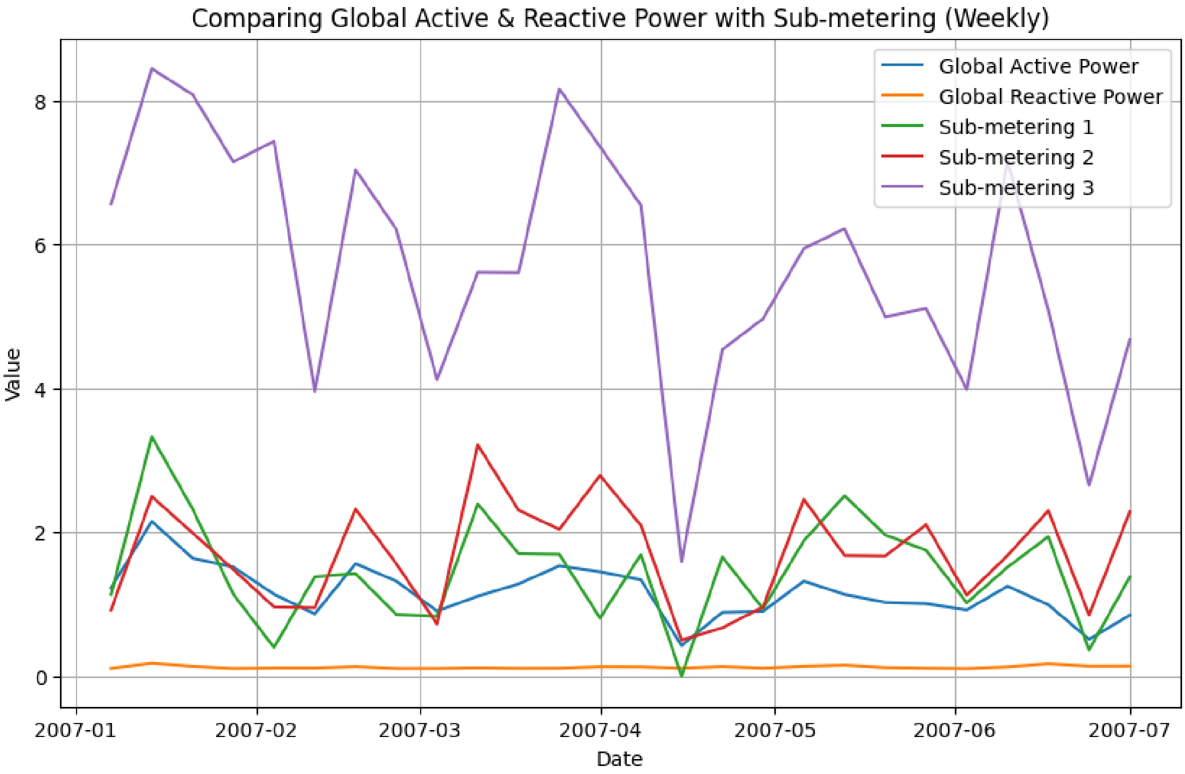
Comparison between sub-meter reading and power consumption monthly.



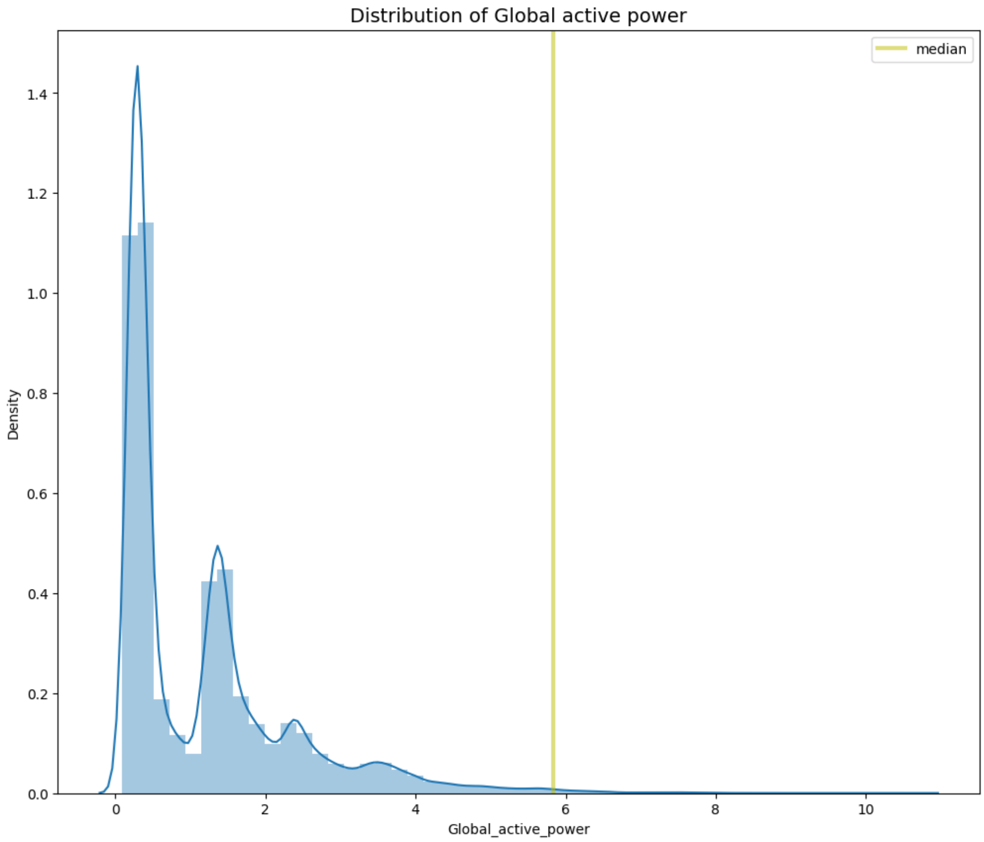
Total energy consumption compared to power consumption weekly



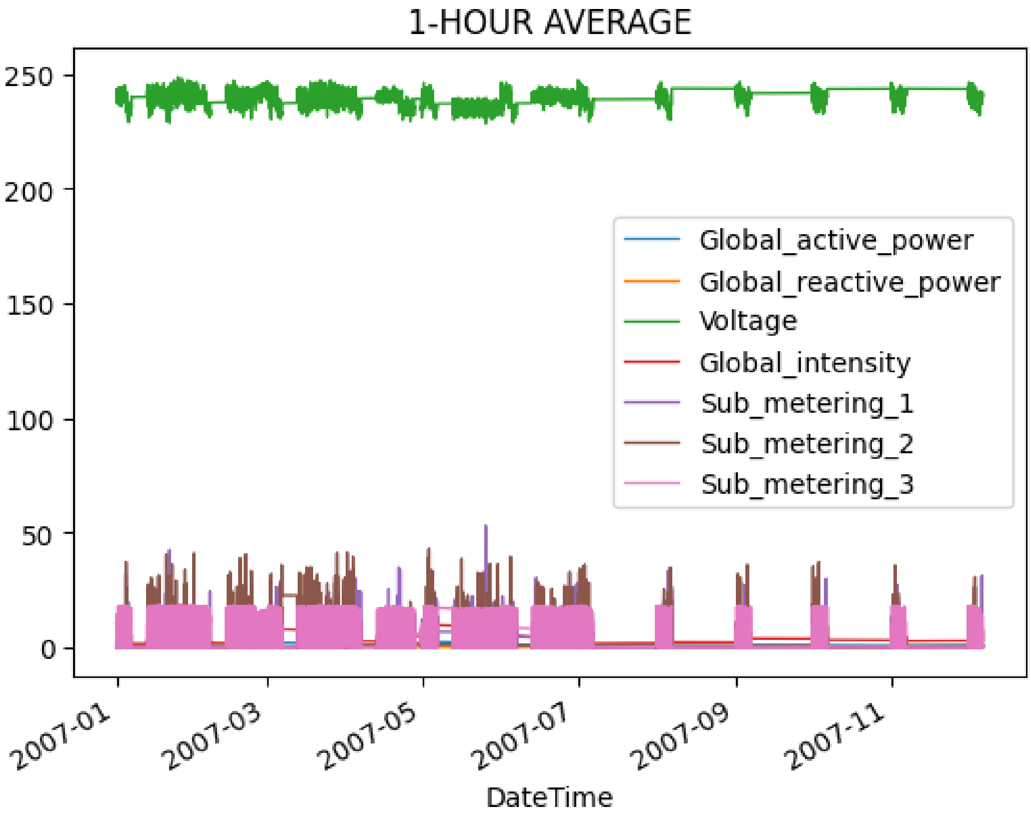
Comparing global active and global reactive with sub-metering weekly.



Distribution of Global active power, showing most reading is lesser then the median.

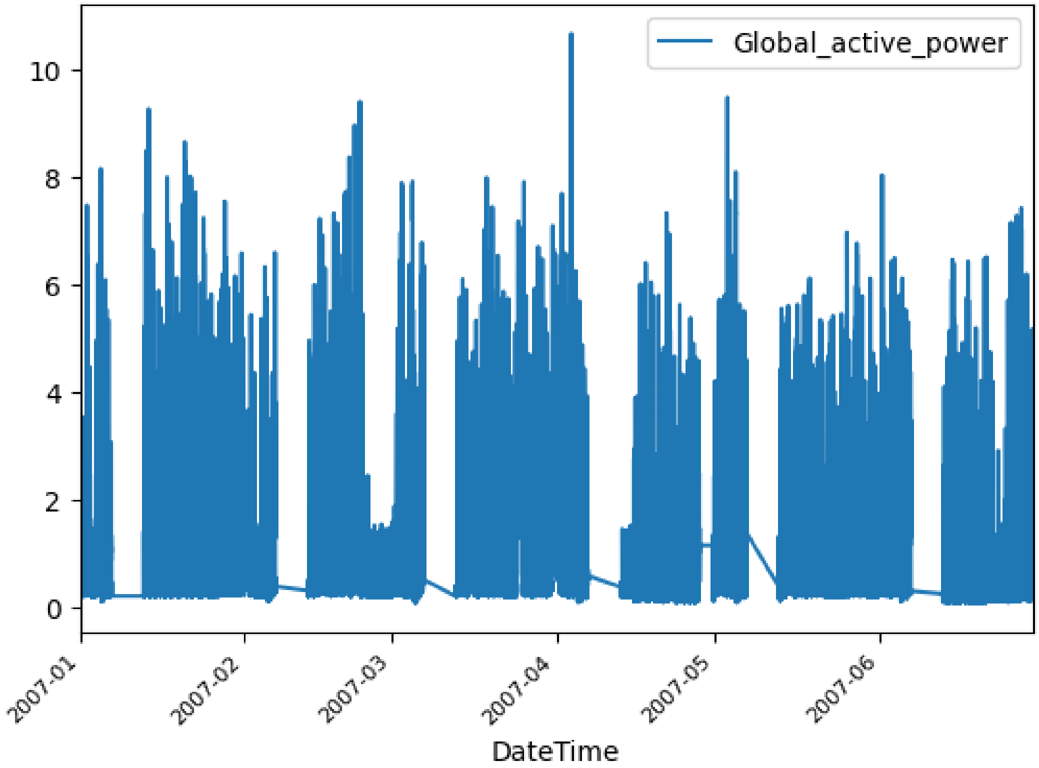


The plot of features showing how voltage is outliers among features.

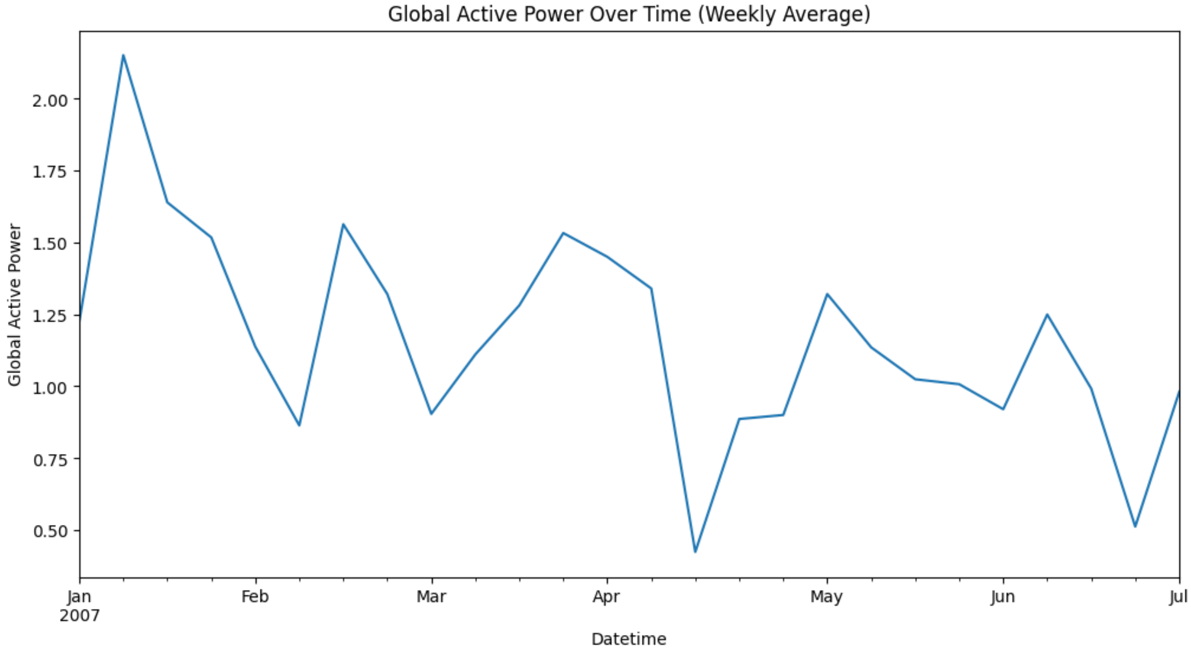


b. ARIMA modelling

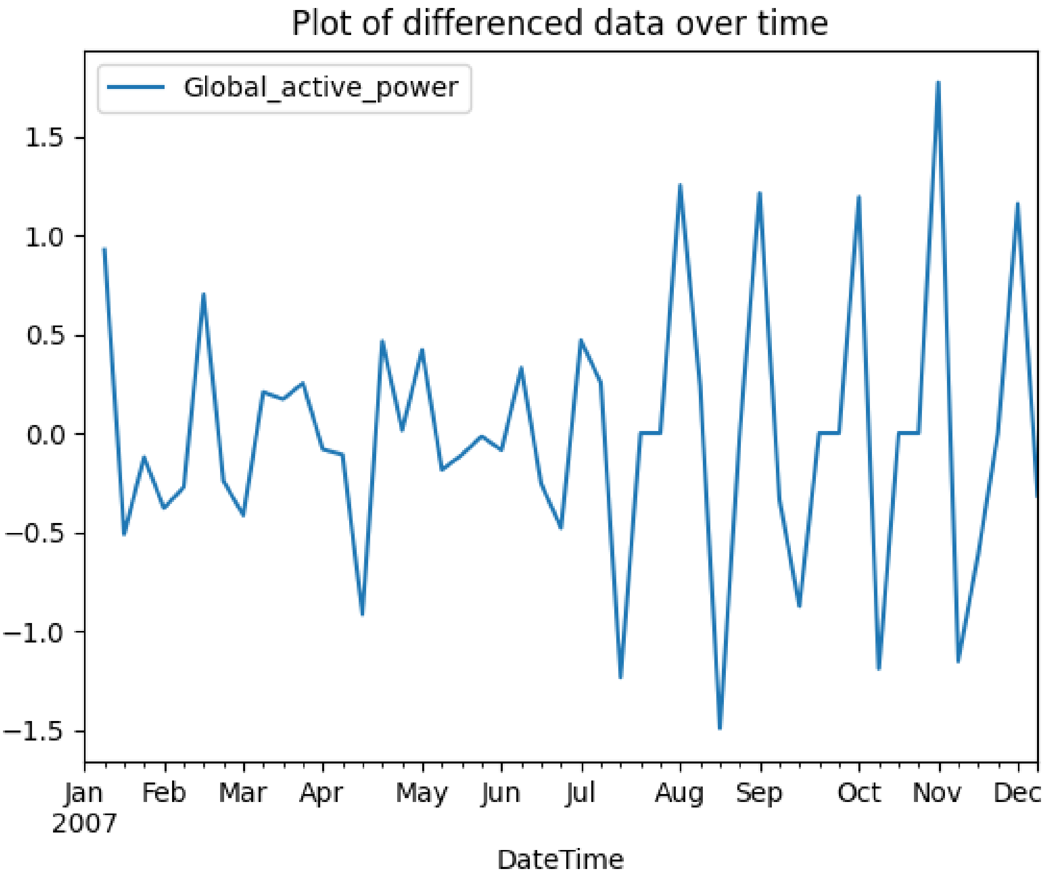
Plotting the original global active power, DateTime as index



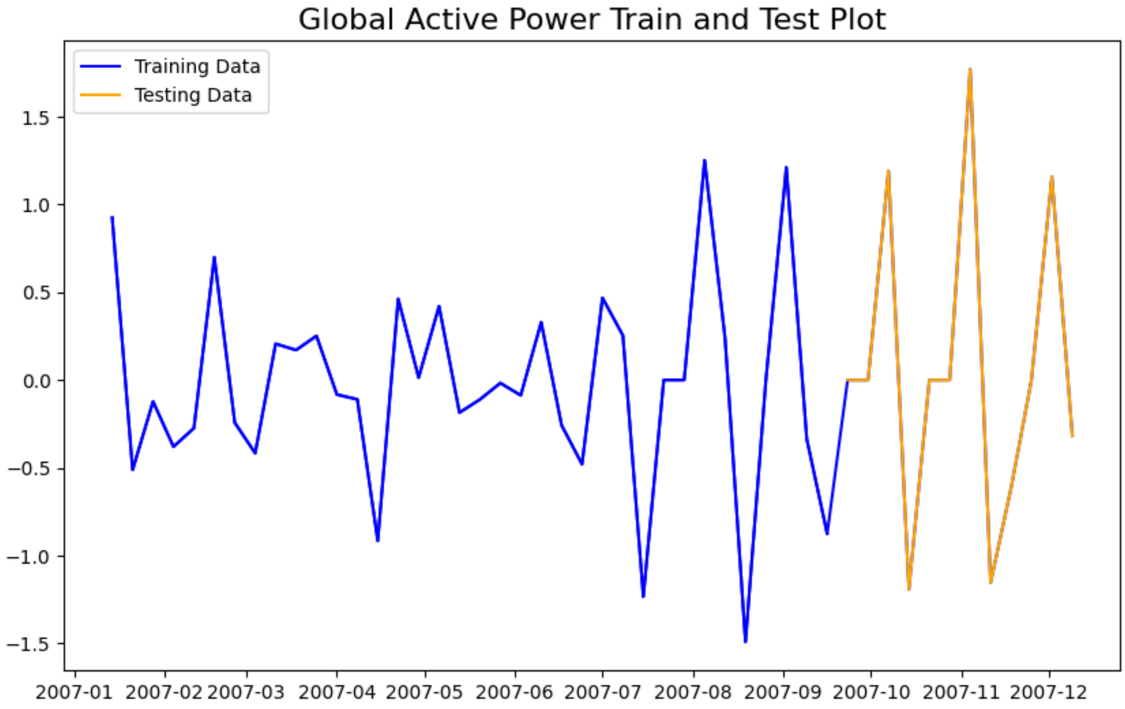
Resample the original data into week, to see the pattern better.



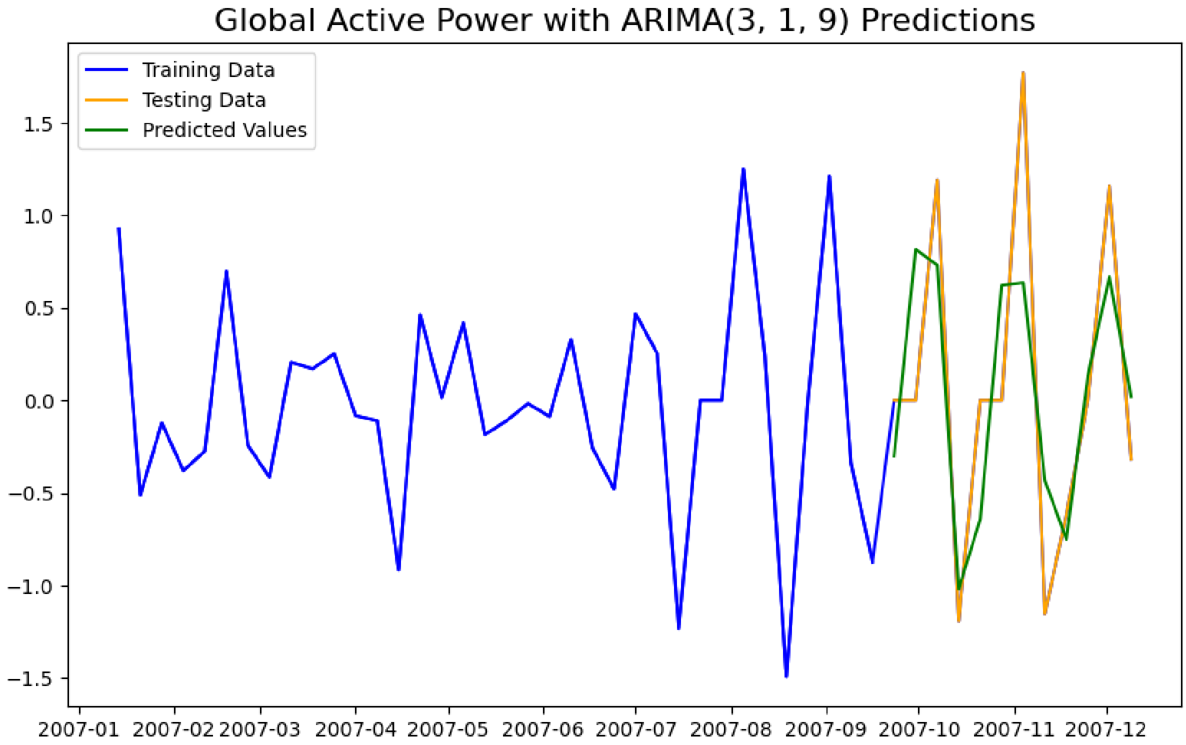
Resampled dataframe is transformed into stationary by differencing.



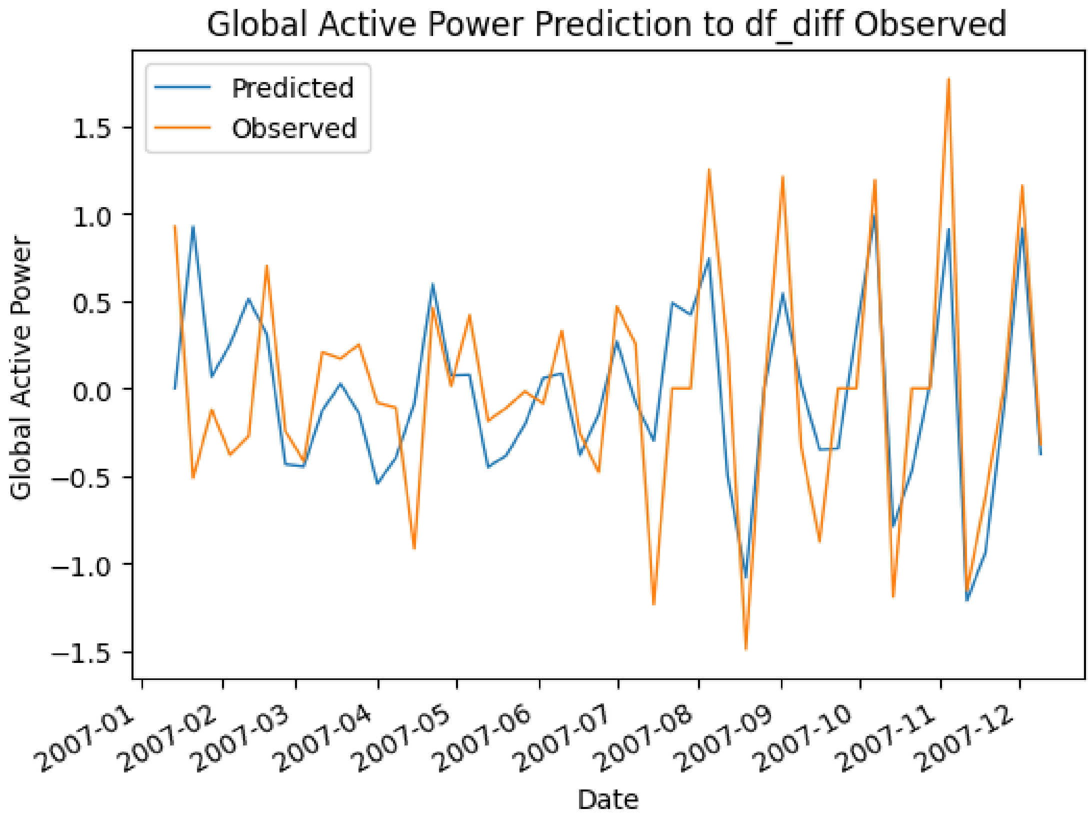
Stationary dataframe is split into (75:25) Train-Test Set



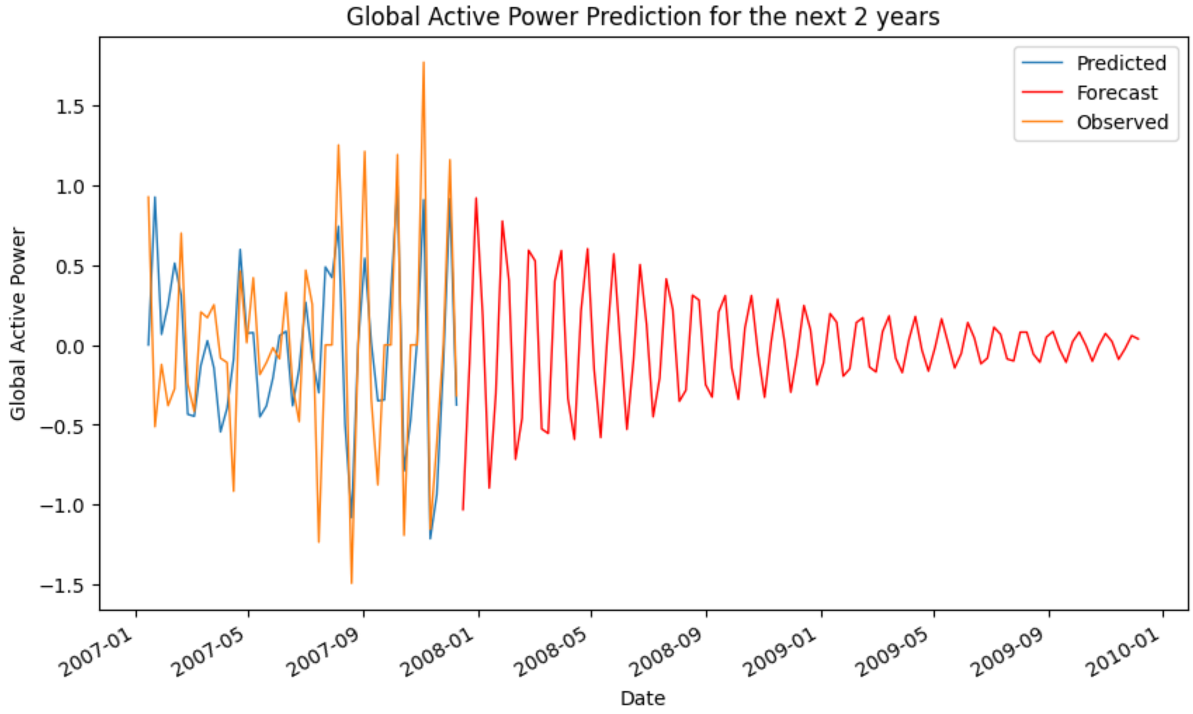
The ARIMA prediction is overlay with Train-Test set.



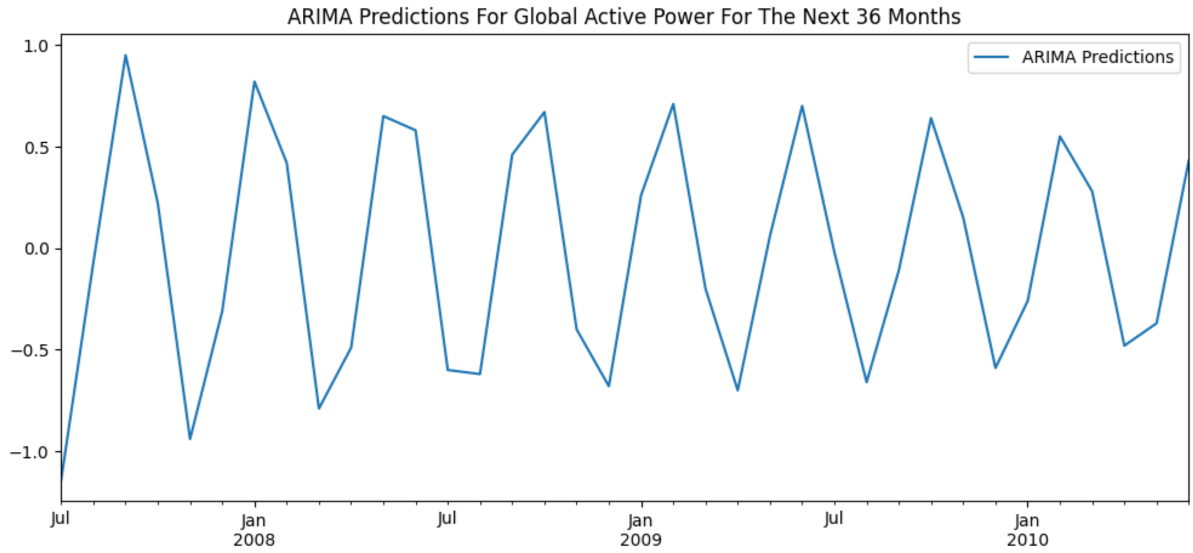
The prediction compared to observed plot.



The next 2 years forecasting from the ARIMA prediction.

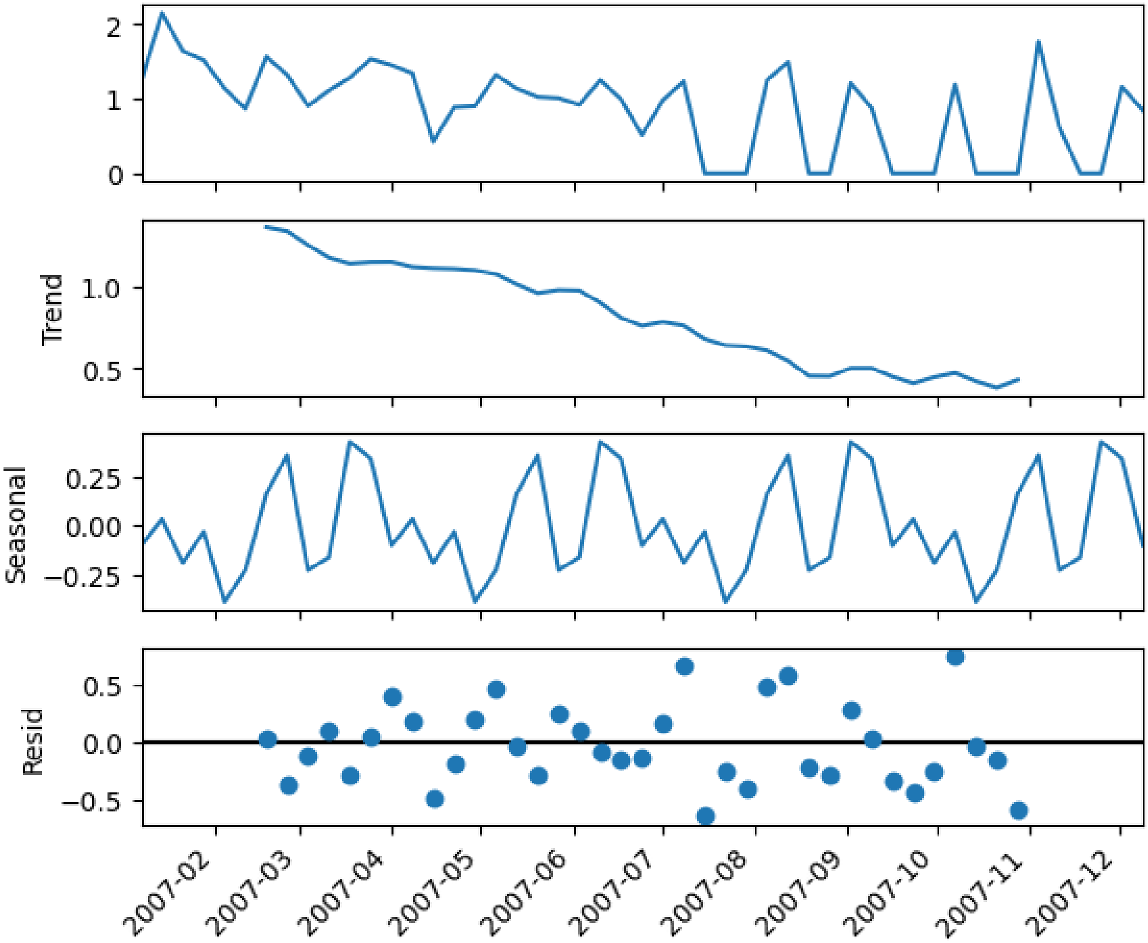


The ARIMA prediction for the next 36 months.

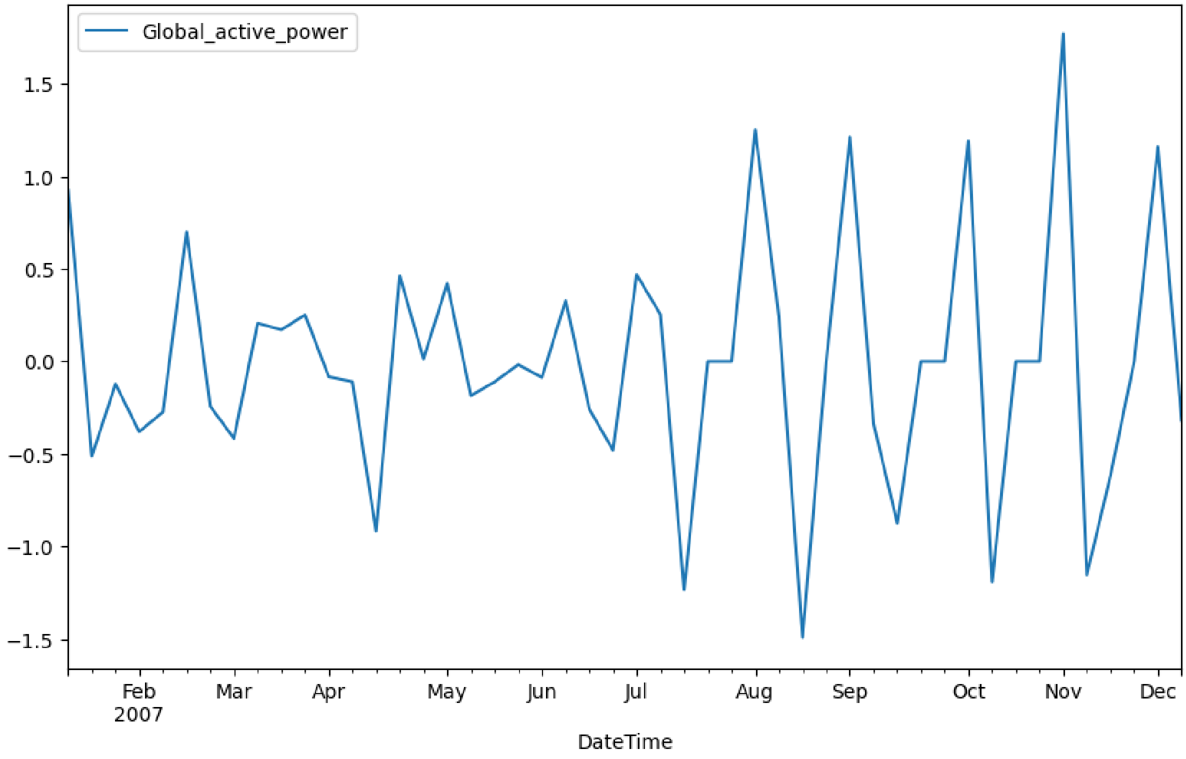


c. SARIMA modelling

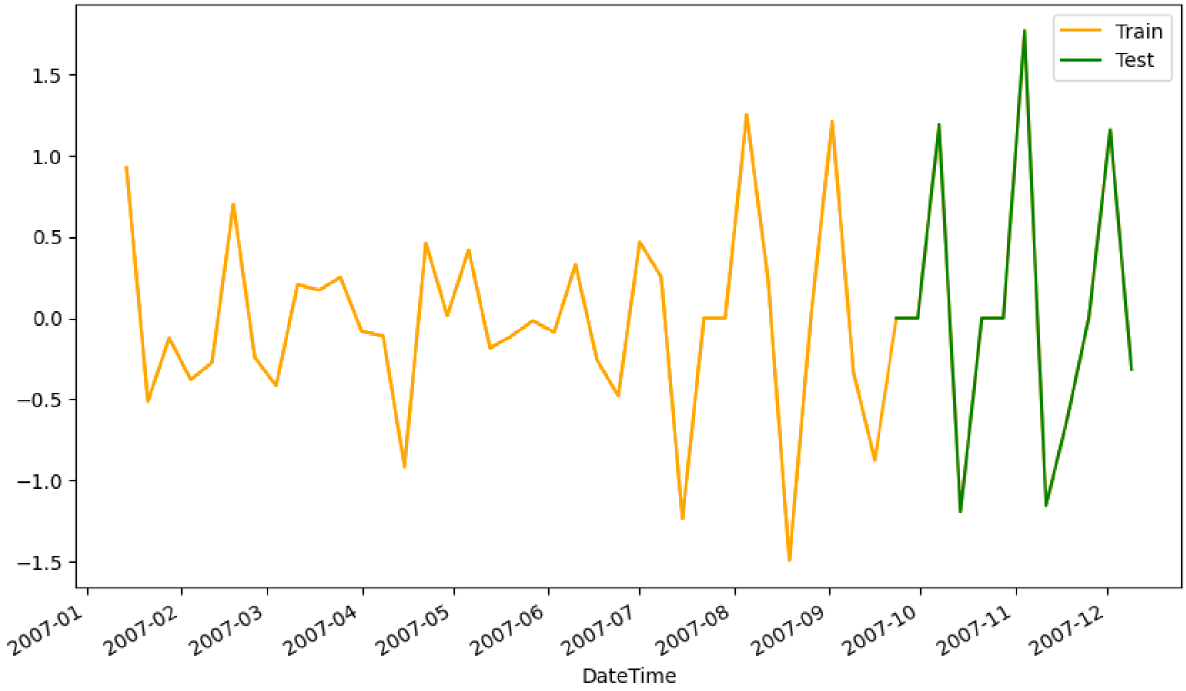
The seasonality for dataset can be seen from seasonal decompose plot



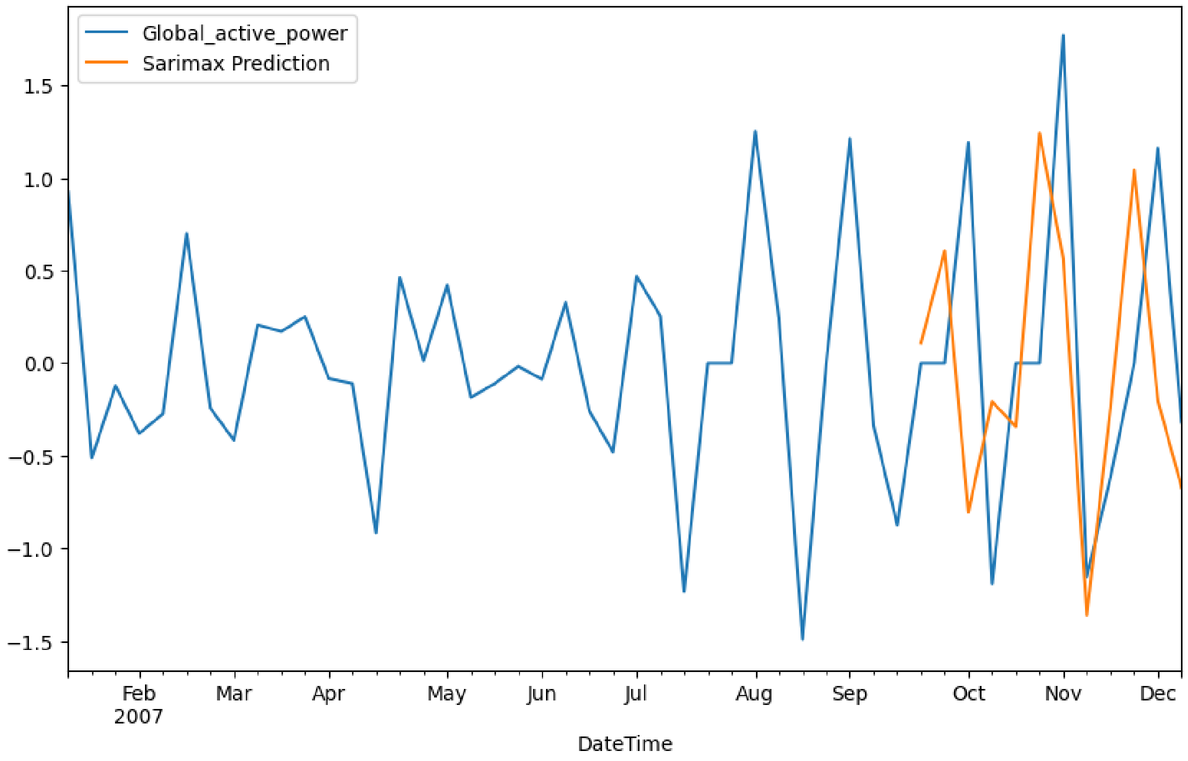
The plot for stationary global active power.



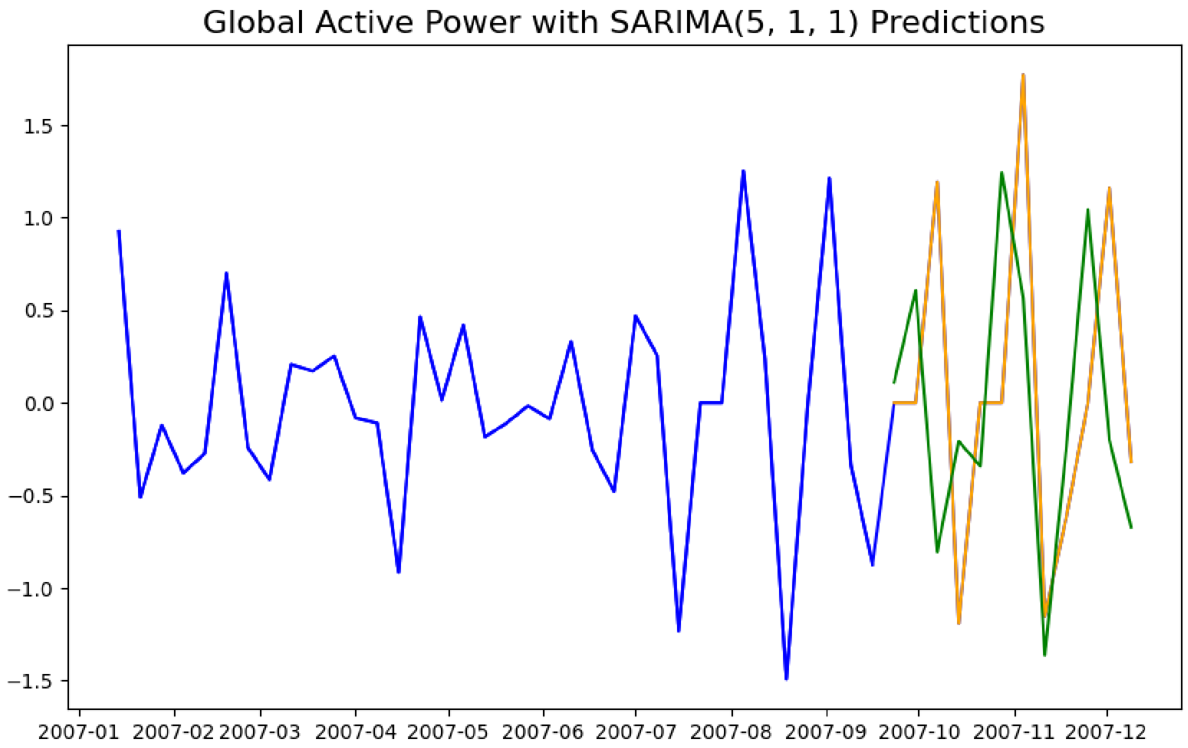
The split plot of train-test set.



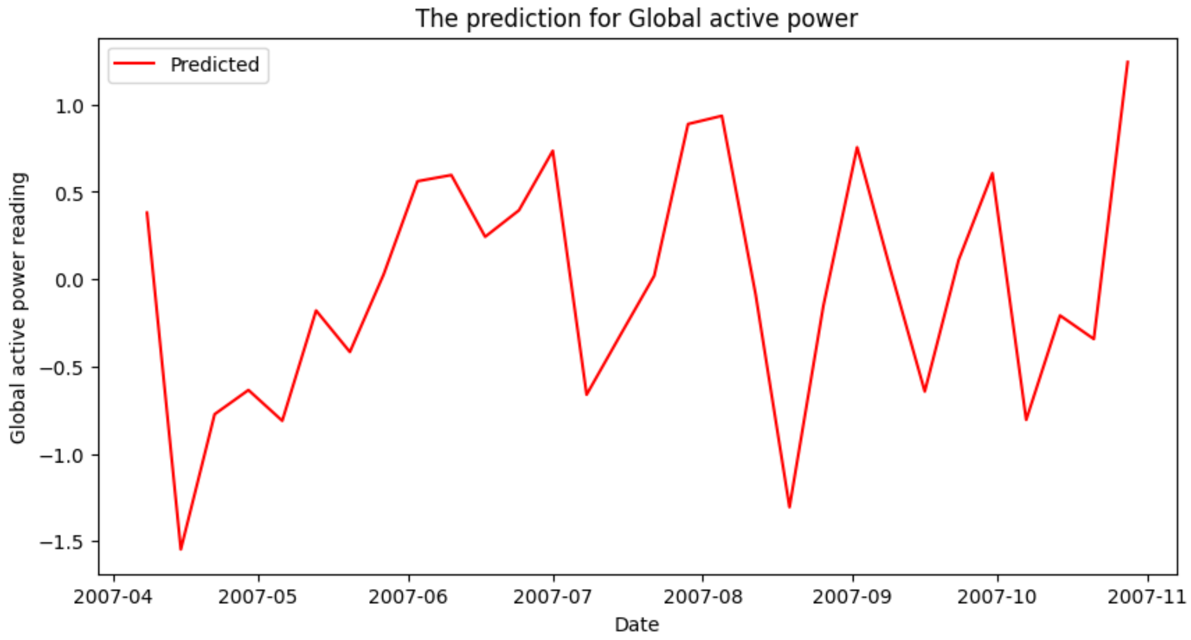
The overlay of SARIMA prediction on the stationary global active power plot.



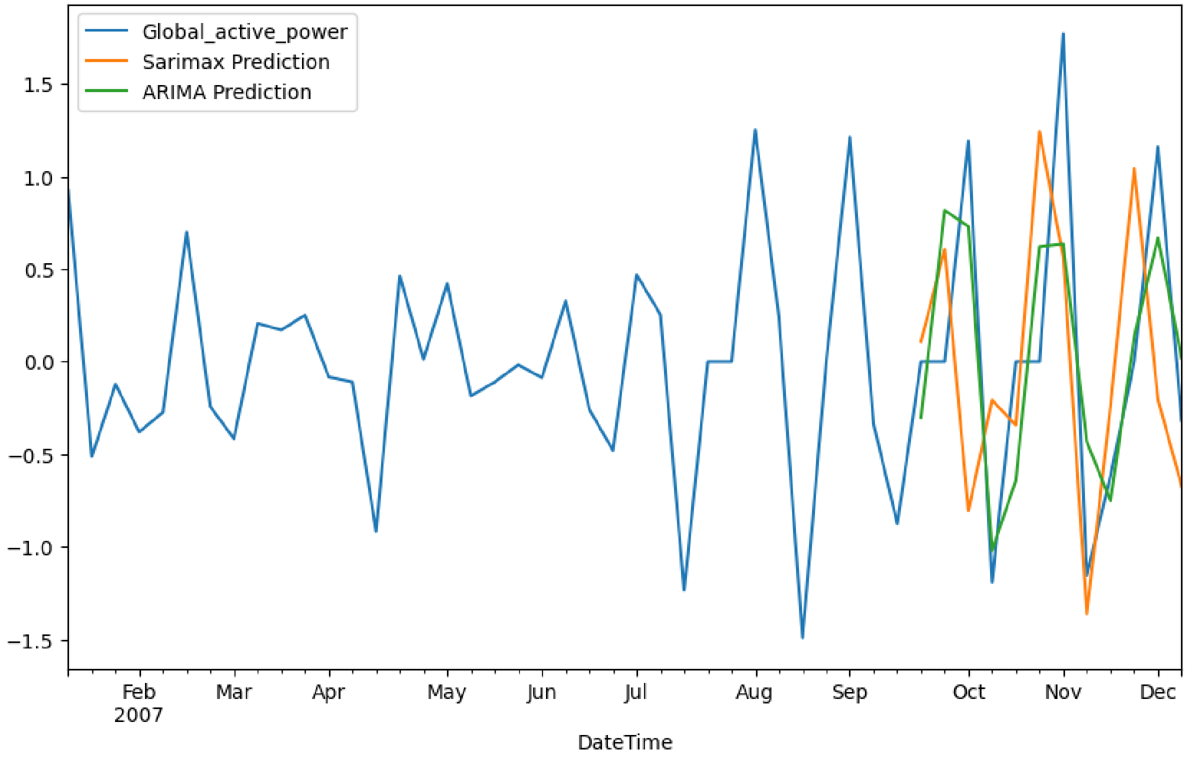
The SARIMA prediction onto the testing set.



The prediction plot for SARIMA model for the next 3 months.



The comparison of ARIMA prediction plot and SARIMA prediction plot on the stationary global active power plot.



3. DOCUMENTATION

What changes on the modelling?	I. High error in ARIMA prediction and seeing the seasonal pattern on seasonal decompose plot of dataset, decide to implement SARIMA modelling. II. Debugging the problem for the high error, using function to find best p, d and q combination based on AIC and RMSE value. III. Changing the way of dataset split into Train-Test set. IV. Changing the resample method.

Note	Remarks
1. Dataset documentation	2/6/2024 (new dataset) 1. Acquire dataset 2. Understanding the dataset 3. Check isnull, 4. Check dtypes 5. Drop all null 6. Handle dtypes 7. Change Date and Time to pd.datetime 8. Proceed to EDA and all 3/6/2024 1. Rather than drop all nulls, fill the nulls with mean 4/6/2024 1. Combine Date and Time column, and formatted to pd.datetime 2. Fill the nulls 3. Address the dtype issue 4. Import as new csv 6/6/2024 (Start modelling) 1. Load csv 2. Datetime column as index 3. Drop all columns except the predictor column 13/6/2024 1. Resample dataset weekly using sum 2. Transform data using differencing 3. Train-Test Split (30 last row as testing set) 24/6/2024 1. Resample dataset weekly using mean 2. Transform data using differencing 3. Fill the null after transform with 0 4. Train-Test Split (75:25 ratio)