Time Series Forecasting Project

Electricity Consumption Forecasting

Guided By:
Monika Bhattacharjee
Assistant Professor, IIT Bombay

Problem Statement

Objective:

Time Series Forecasting of Household Electricity Consumption using ARIMA and SARIMA Models

Emphasis:

Model diagnostics and validation to ensure reliable forecasting of electricity consumption and to assess how well ARIMA and SARIMA models capture trends and seasonality in the data

About the Dataset

The dataset, sourced from UCI Machine Learning Repository, is having **20,75259 rows** , and 9 columns and we extracted following three columns for our further analysis.

Columns:

Column Name	Description
Date	The calendar date of observation, recorded in the format dd/mm/yyyy (e.g., 16/12/2006).
Time	The specific time of day in hh:mm:ss format (e.g., 17:24:00). It is combined with Date to form a timestamp for each observation.
Global Active Power	The total active power consumed in the household at a given time (in kilowatts). It represents the real power that performs useful work (e.g., running appliances).

Glimpse of the Data

	Date	Time	Global_active_power
0	16/12/2006	17:24:00	4.216
1	16/12/2006	17:25:00	5.360
2	16/12/2006	17:26:00	5.374
3	16/12/2006	17:27:00	5.388
4	16/12/2006	17:28:00	3.666

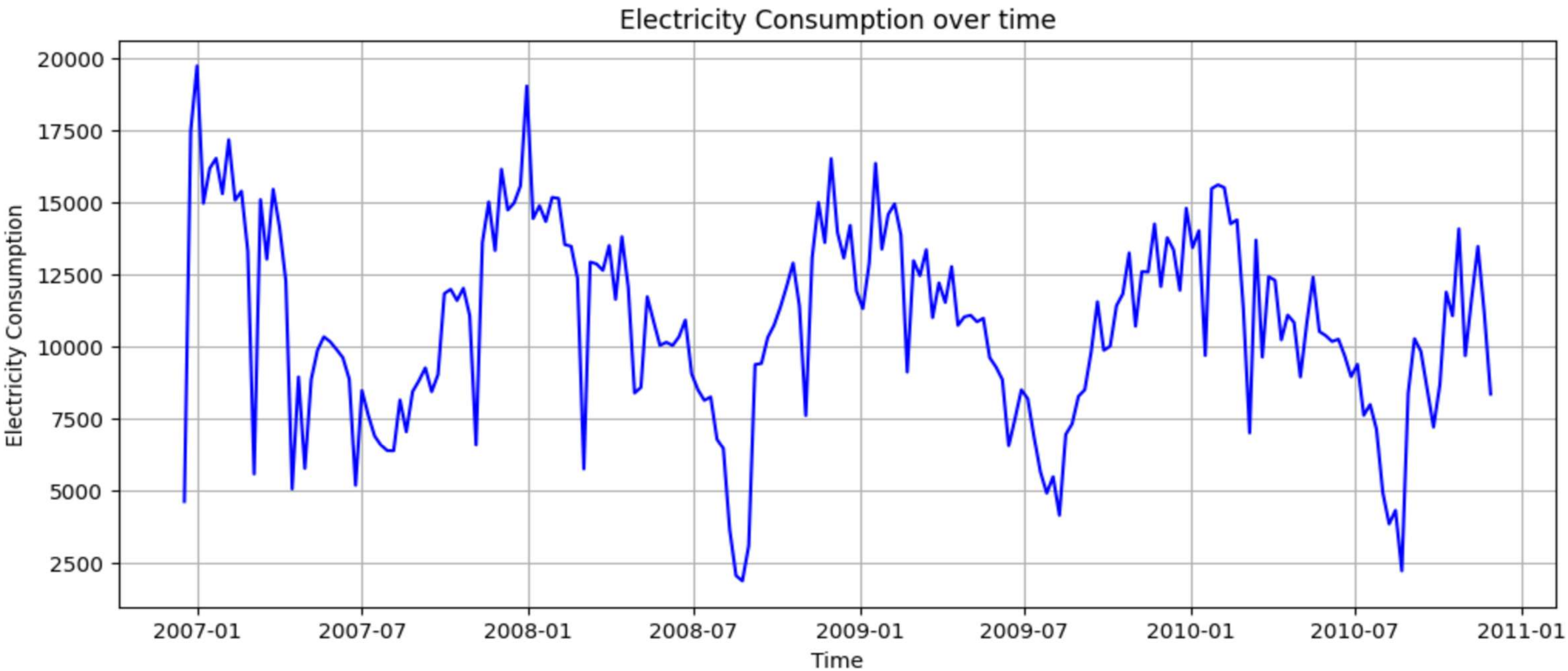
Preprocessing of the data

- There are 1.25% missing values in the minute dataset.
- Assembled the columns Date and Time to make a new column as DateTime.
- Since missing values are less than 5%, hence we removed it and further resampled it in weekly data.
- Assigned Datetime column as index of the series "Global_active_power".

	Global_active_power
Datetime	
2006-12-17	4599.636
2006-12-24	17477.128
2006-12-31	19736.518

Plot of Electricity Consumption Over Time

- ❑ **Series seems stationary in nature & also reflecting seasonal bahaviour.**



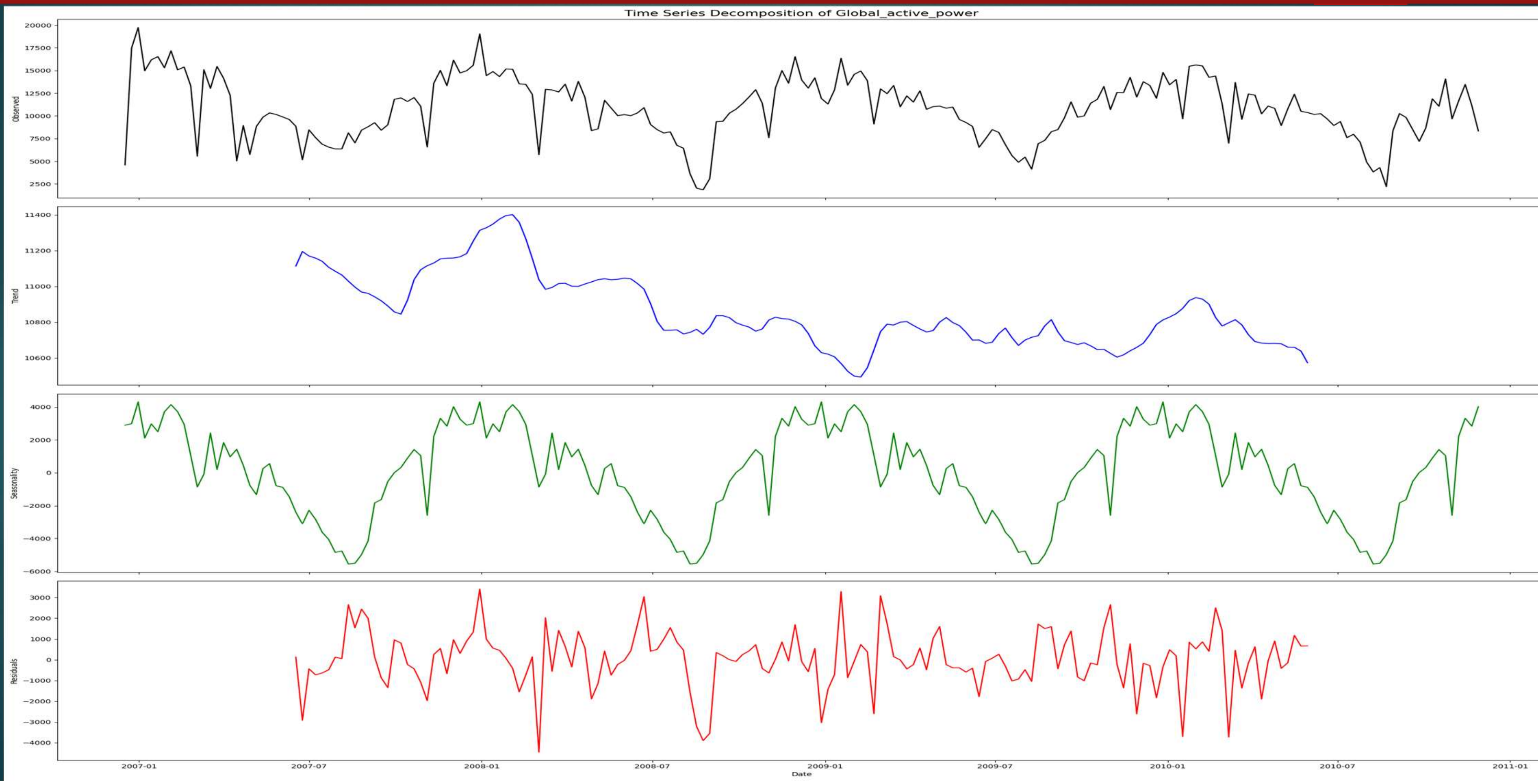
Decomposition of the Series

Observed

Trend

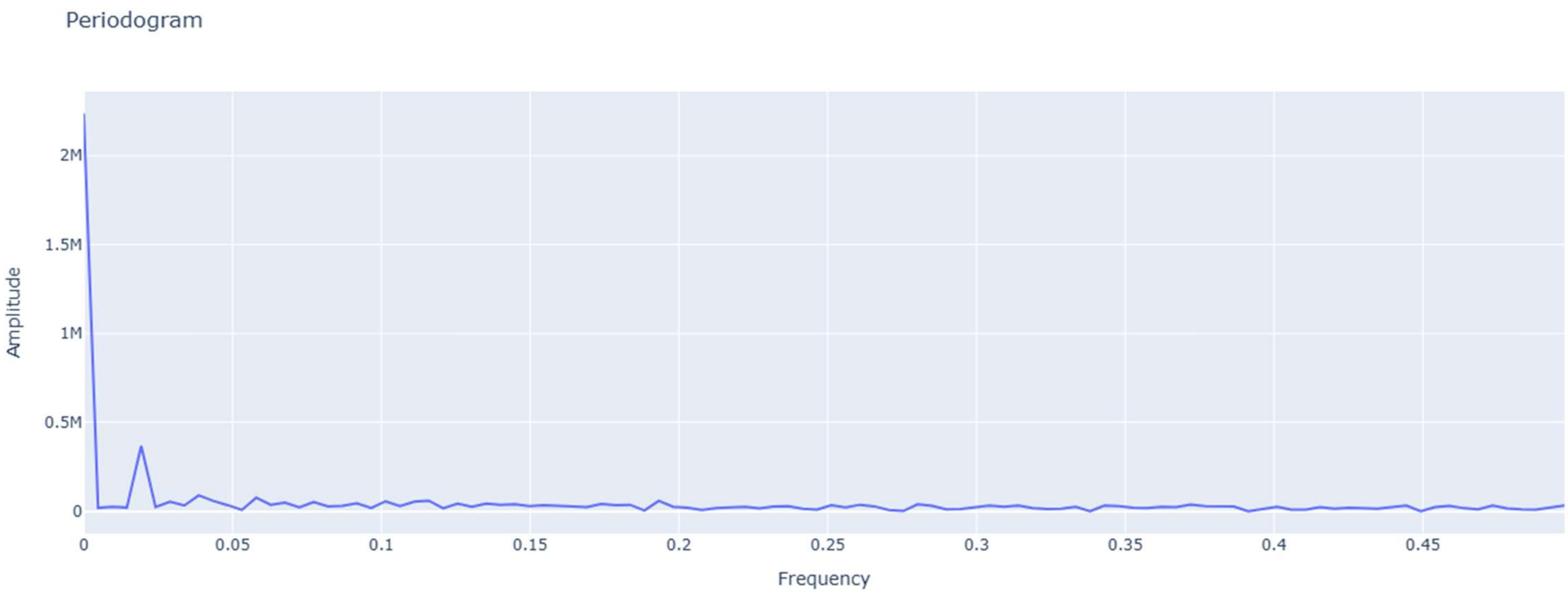
Seasonality

Residual



Testing for Seasonality

- We found a highest spike at $x=0.01932367$ which is nothing but 1/52
- Hence, here in the weekly data yearly seasonality is present with period=52 weeks



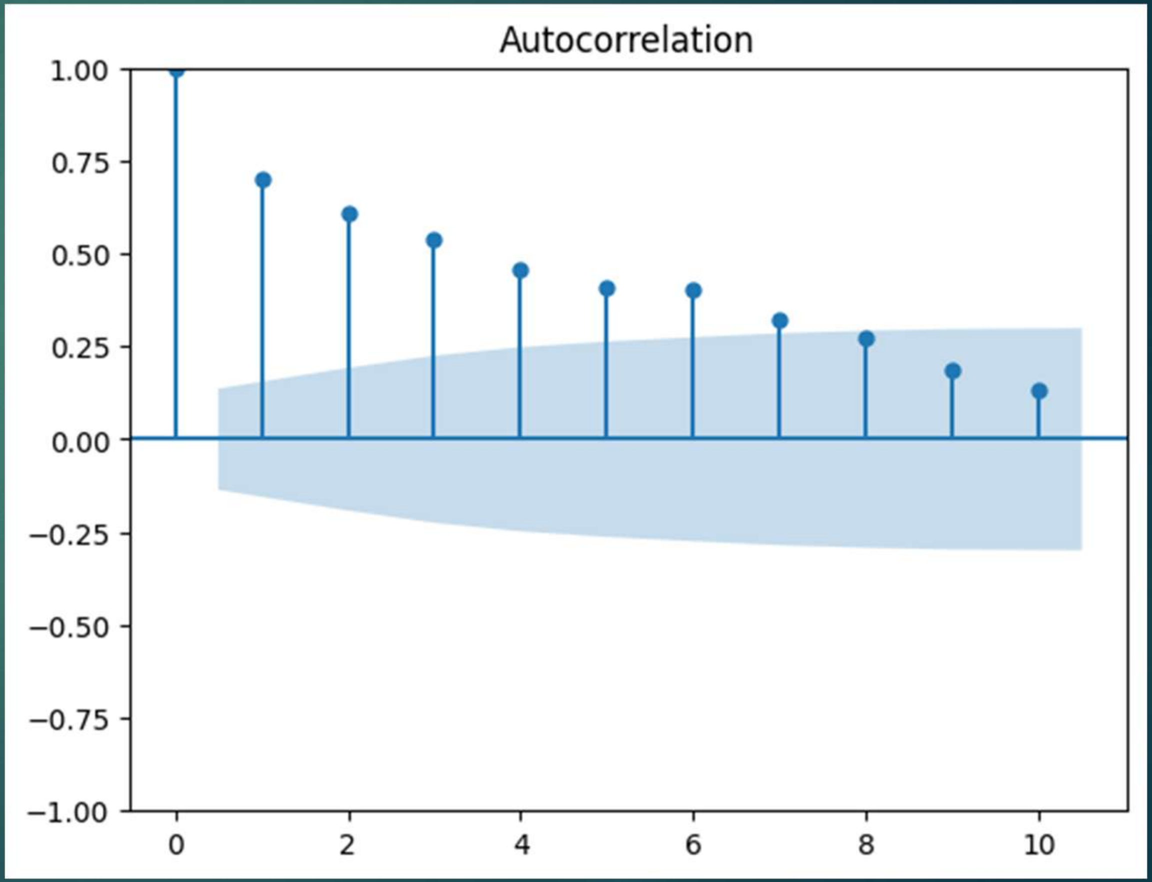
ADF Test For Stationarity

- ADF Test Results
  - Test statistic: *-4.302692*
  - p-value: *0.000439*

- Hence, *p-value < 0.05,* the time series is stationary

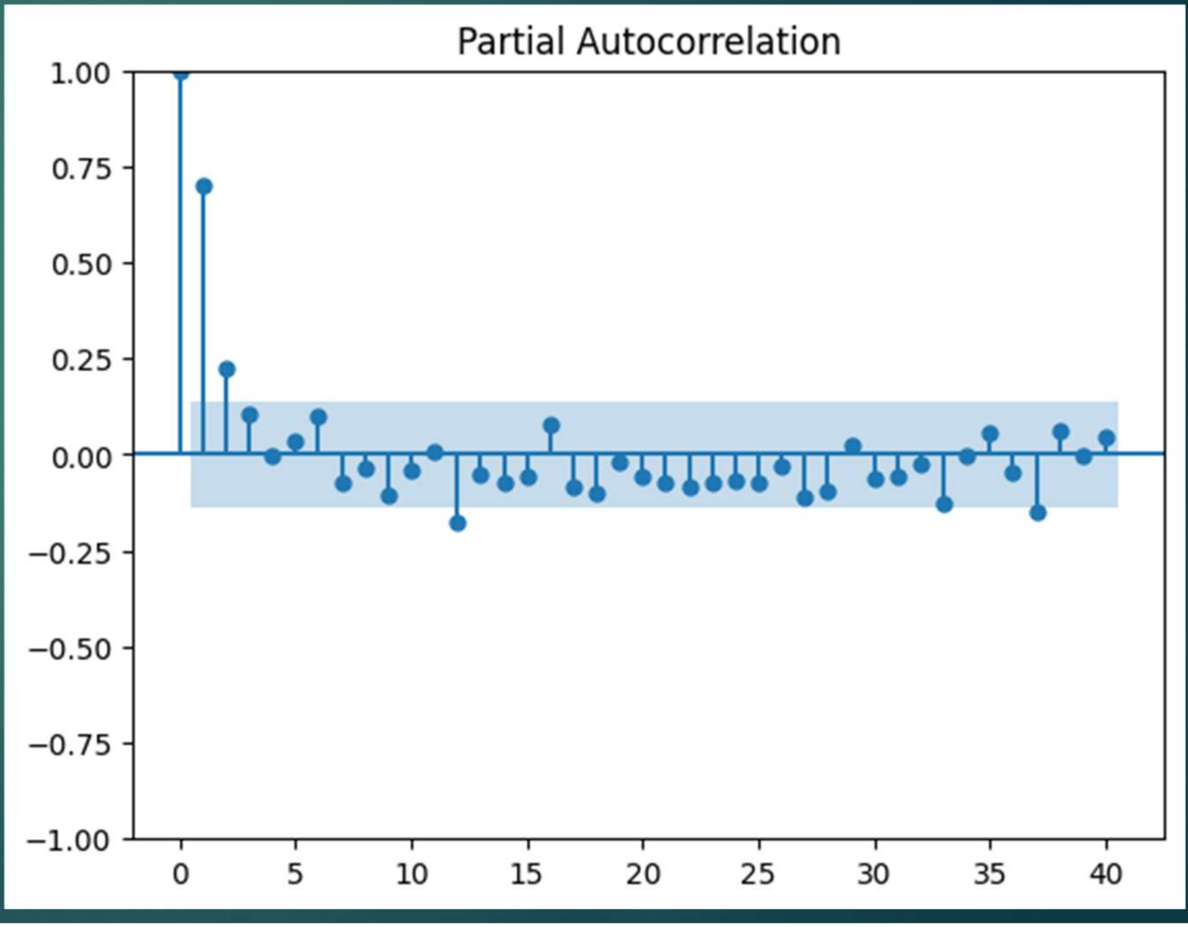
- This shows exponential decay so we should go for PACF plot

ACF Plot



PACF Plot

- ❑ PACF plot indicates that this series might be AR(2) process
- ❑ Now let's split the data in Train and Test set
- ❑ We will keep 95% data for train set and 5% for test set
- ❑ Stepping towards model building….



ARIMA Model

- ❑ Fitting various ARIMA model for different (p, q) values
- ❑ Since the series is stationary, hence parameter *d=0*
- ❑ ARIMA (2, 0, 5) model has the least AIC value
- ❑ Thus, we concluded ARIMA (2, 0, 5) is the final ARIMA model

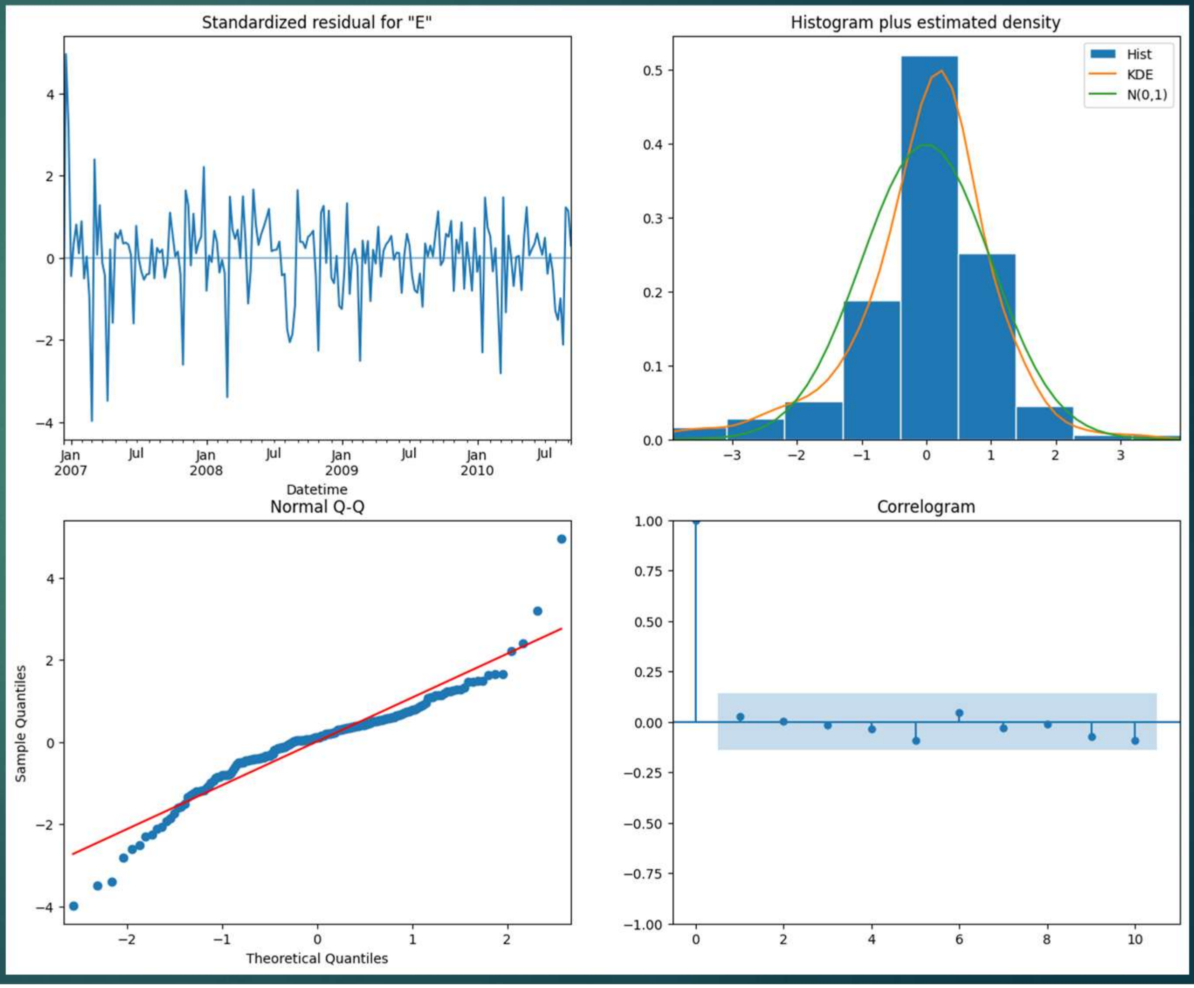
	Model	AIC
17	[(2, 0, 5)]	3580.623797
23	[(3, 0, 5)]	3585.589613
21	[(3, 0, 3)]	3600.296919
27	[(4, 0, 3)]	3601.521034
22	[(3, 0, 4)]	3601.931163

Model Diagnostics ARIMA (2, 0, 5)

❑ **Remark**:

- According to the qualitative analysis, residuals approximately resemble white noise.

- In order to validate this, let’s shift to quantitative analysis, namely Ljung-Box test



Ljung Box test

- **Remark**:
  - We have checked Ljung Box test upto 10 lags
  - p-values indicates that residuals are normal & not autocorrelated
  - We can move towards forecasting using ARIMA (2, 0, 5)…..

	lb_stat	lb_pvalue
1	0.002163	0.962907
2	0.052353	0.974163
3	0.088128	0.993223
4	0.314838	0.988836
5	2.062428	0.840441
6	2.612314	0.855692
7	2.845476	0.898920
8	2.848122	0.943524
9	4.236804	0.895147
10	5.878087	0.825400

Forecasting ARIMA(2,0,5)

❑ The MAPE value obtained for the forecast is 17.82%

SARIMA Model

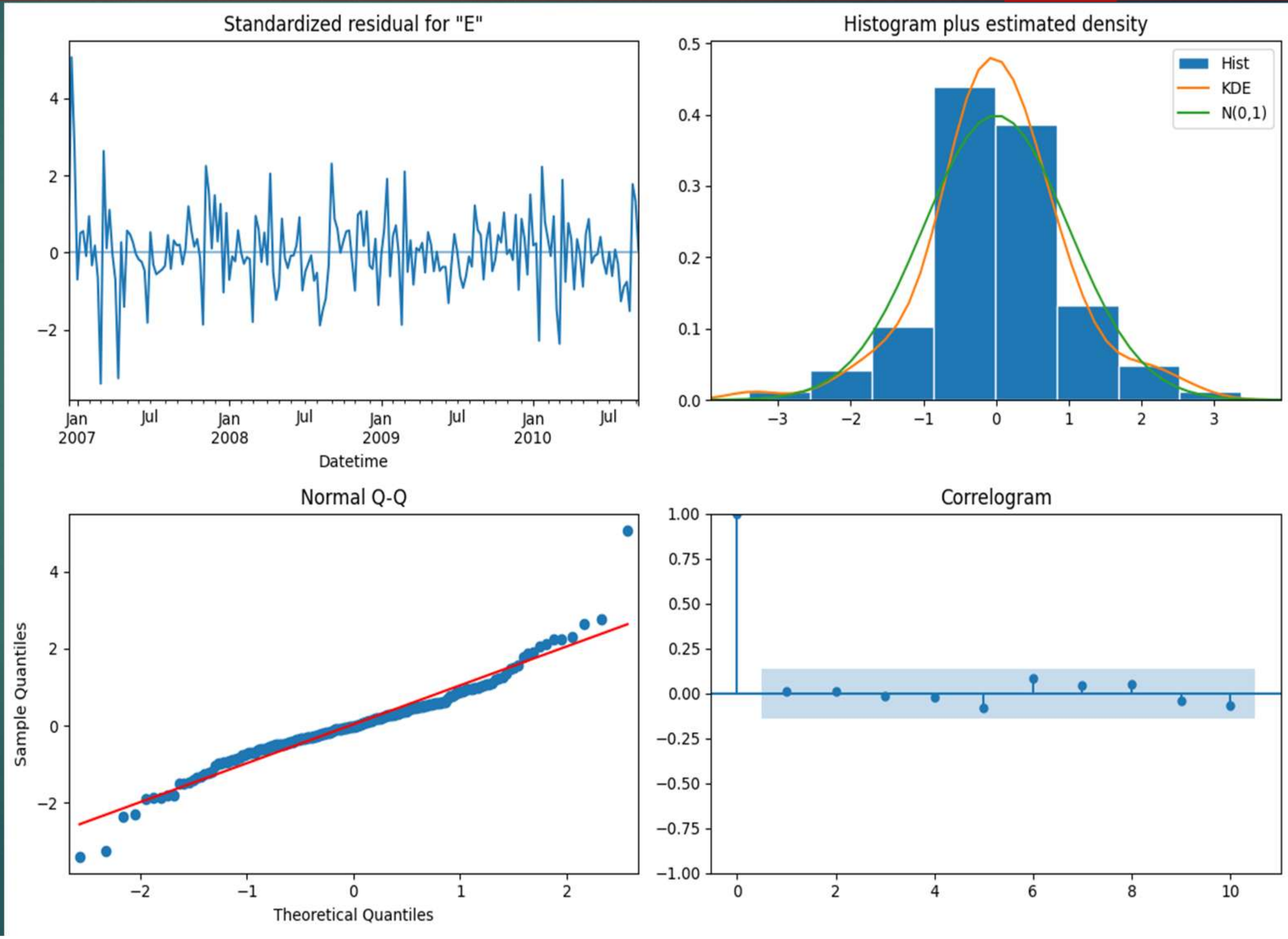
- To further improve the forecasting accuracy/MAPE, we apply SARIMA model as well
- Since the series is seasonal in nature with a period of 52 weeks, so it is suspected that SARIMA model would out-perform ARIMA model
- SARIMA $(2,0,2)(1,0,0)_{52}$ is the best model so far as its AIC value is lowest among all other

	Model	AIC	MAPE
24	[(2, 0, 2), (1, 0, 0)]	3600.851061	0.154902
20	[(2, 0, 1), (1, 0, 0)]	3602.667495	0.169335
16	[(1, 0, 1), (1, 0, 1)]	3602.790993	0.162705
29	[(1, 0, 1), (1, 0, 0)]	3602.799110	0.164127
14	[(1, 0, 2), (1, 0, 0)]	3603.007252	0.165691
10	[(2, 0, 1), (1, 0, 1)]	3603.142120	0.165328
1	[(1, 0, 2), (1, 0, 1)]	3603.468050	0.164001
35	[(1, 0, 1), (0, 0, 1)]	3605.721377	0.174948
30	[(2, 0, 1), (0, 0, 1)]	3606.511345	0.189689
23	[(1, 0, 2), (0, 0, 1)]	3606.729665	0.182354

Model Diagnostics: SARIMA $(2,0,2)(1,0,0)_{52}$

Remark:

- According to the qualitative analysis, residuals approximately resemble white noise.
- In order to validate this, let's shift to quantitative analysis, namely Ljung-Box test



Ljung Box test - SARIMA

- **Remark**:
  - We have checked Ljung Box test up-to 10 lags
  - p-values indicates that residuals are normal & not autocorrelated
  - We can finally move towards forecasting using SARIMA $(2,0,2)(1,0,0)_{52}$ .....

	lb_stat	lb_pvalue
1	0.812376	0.367419
2	1.020928	0.600217
3	1.064478	0.785656
4	1.135710	0.888566
5	2.086840	0.837003
6	3.567310	0.734994
7	4.074613	0.771145
8	4.636405	0.795636
9	5.080507	0.827234
10	6.255841	0.793330

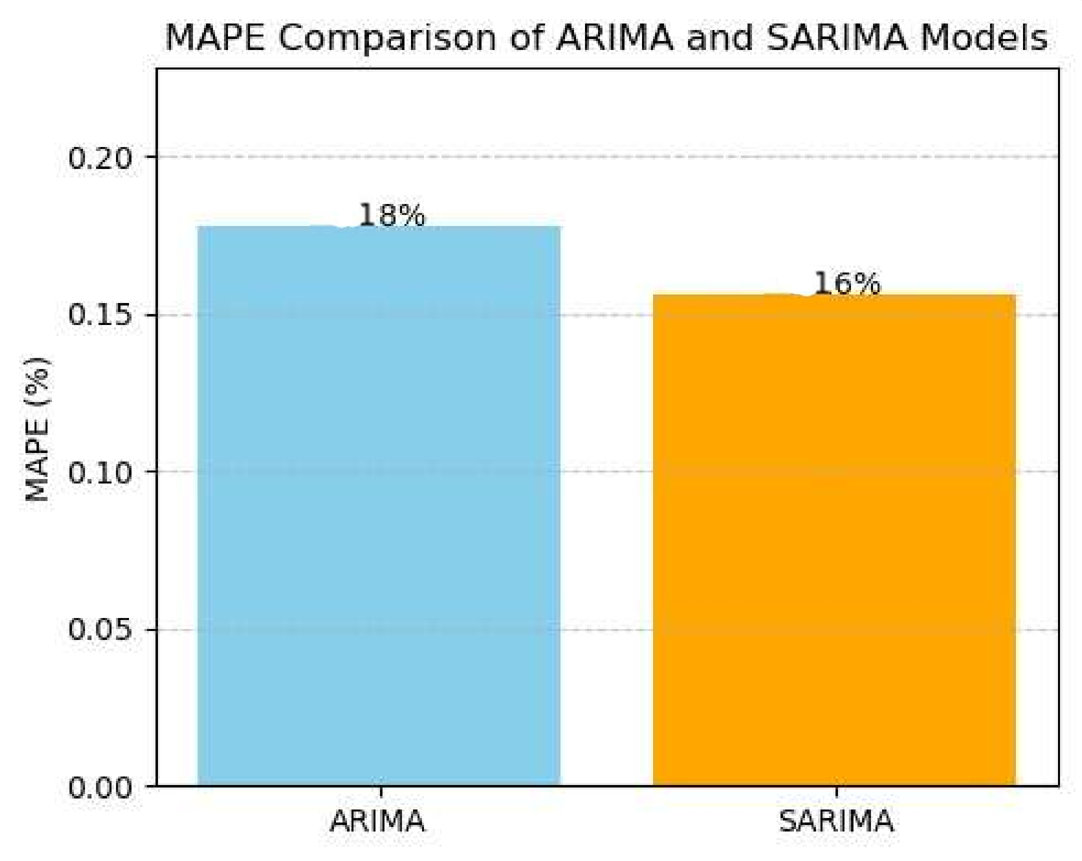
Forecasting - SARIMA

- The MAPE value obtained for the forecast is 15.49%

Comparison: ARIMA vs SARIMA

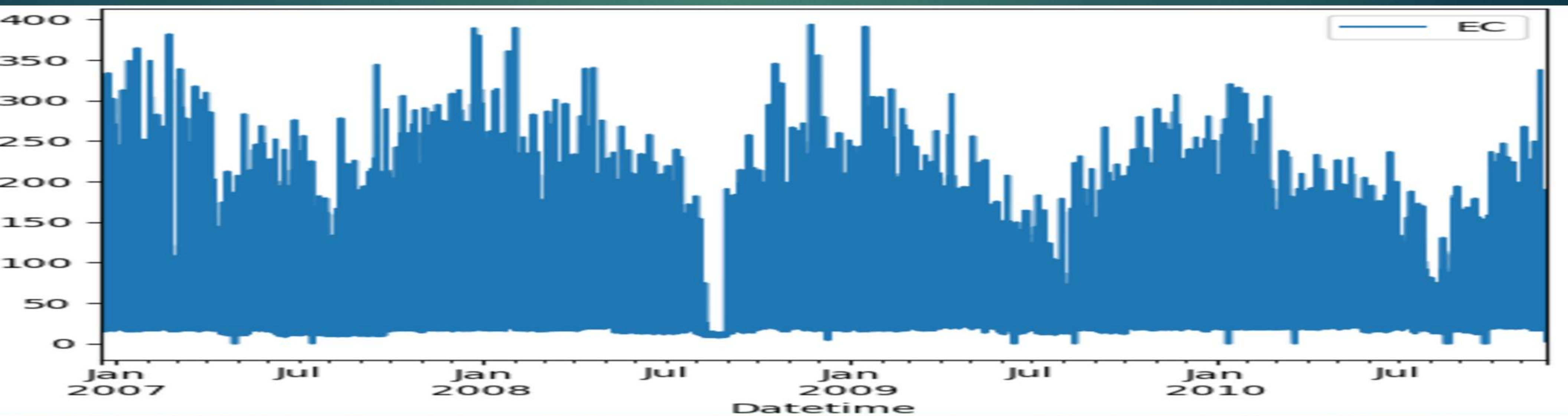
❑ **Remark**:

- SARIMA model has outperformed the baseline model ARIMA
- Making the seasonality a distinguishing factor



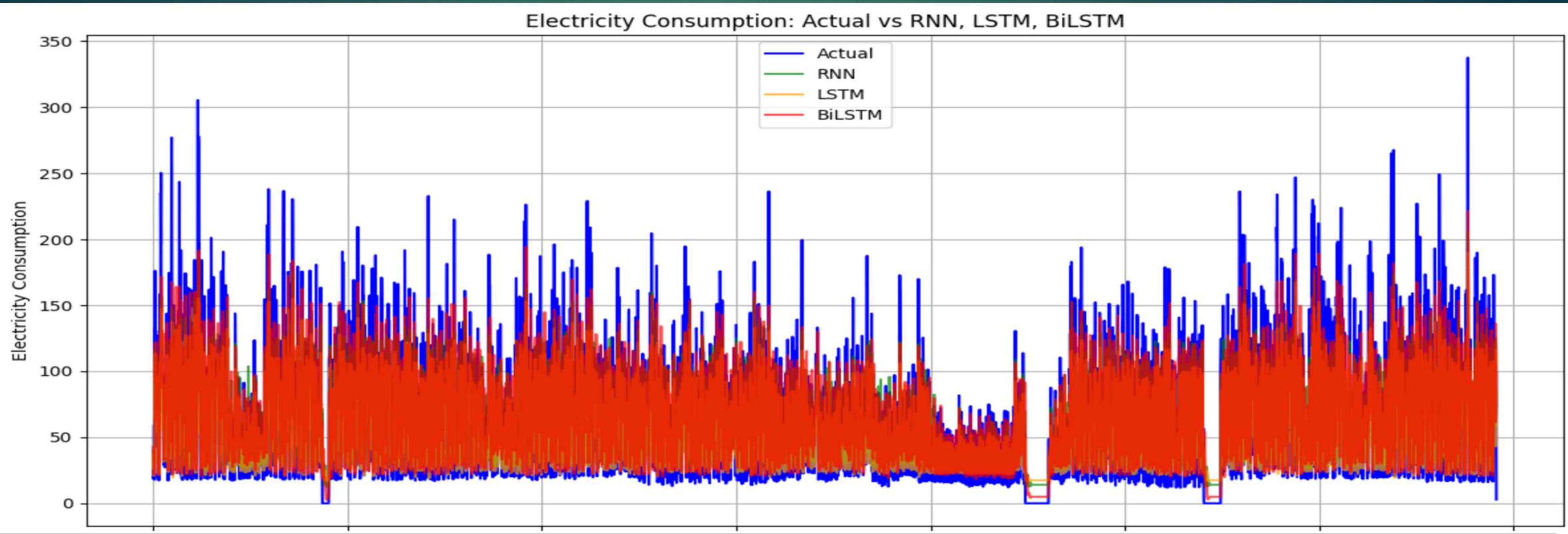
Towards advanced models

- Resampling the original dataset with 20 lacks+ rows to hourly
- The resampled data has 34579 rows so we can use “Deep Learning Models as well.”
- Used Minmax scalar for the scaling the series since DL models perform better on the scaled data



RNN,LSTM & BiLSTM Models

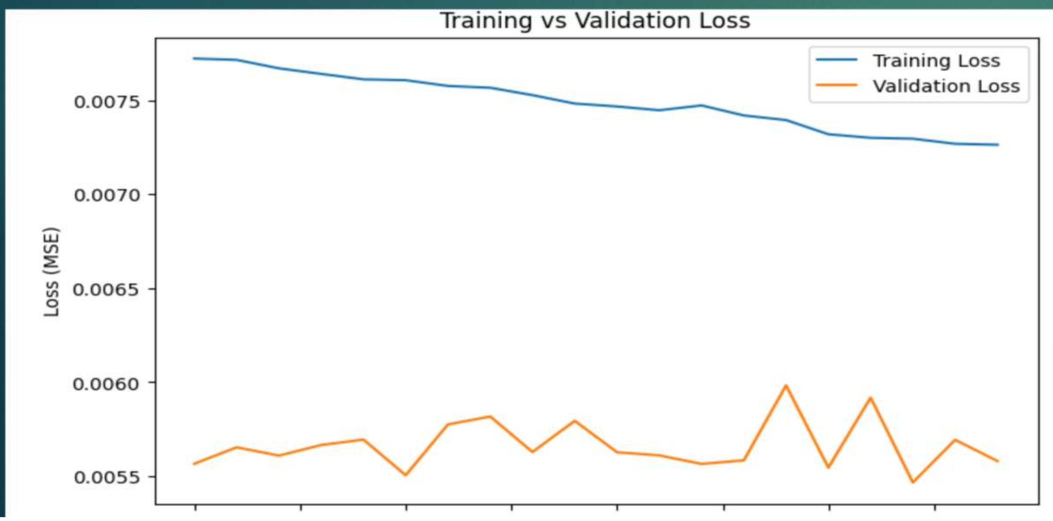
- Fitted RNN, LSTM and BiLSTM deep learning models on the hourly dataset
- Calculated the MAPE values corresponding to each model to compare them

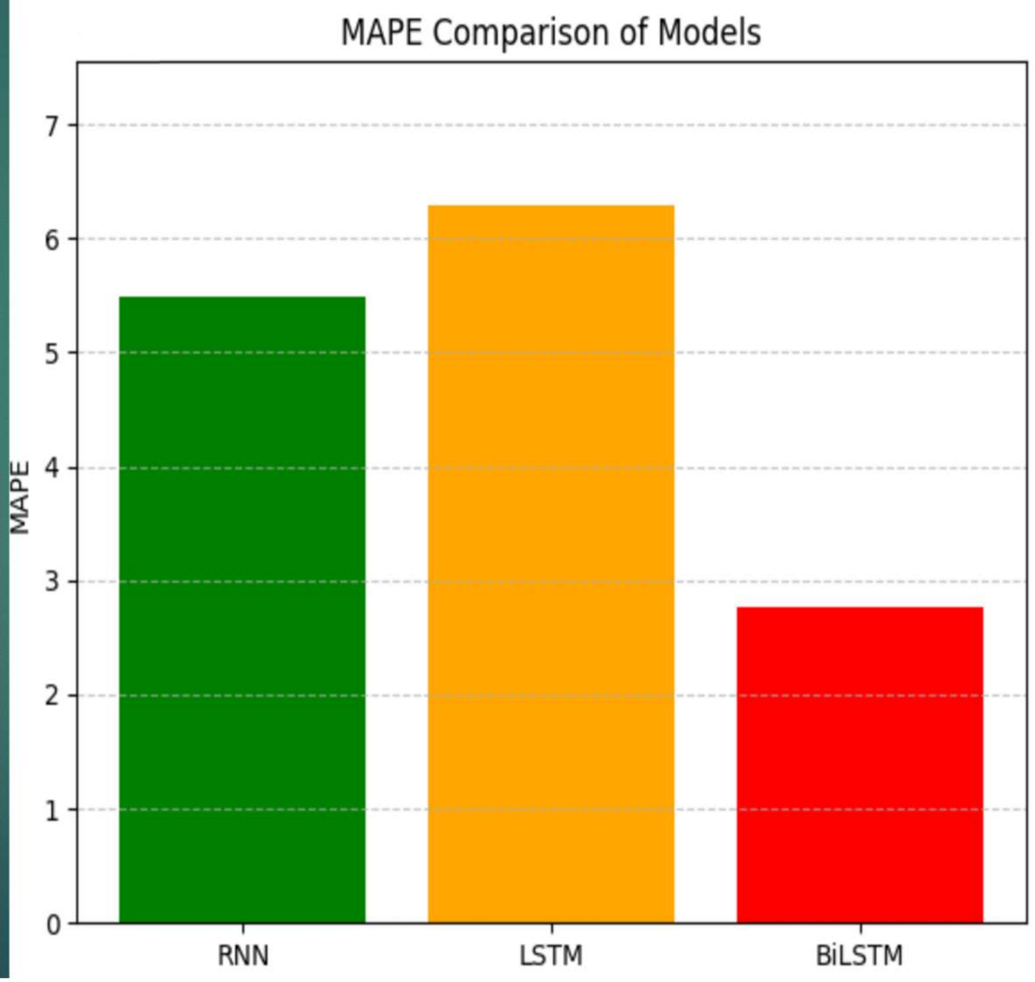


Camparison of DL Models

- As the MAPE values suggest that the BiLSTM model outperformed RNN & LSTM with MAPE score 2.9%

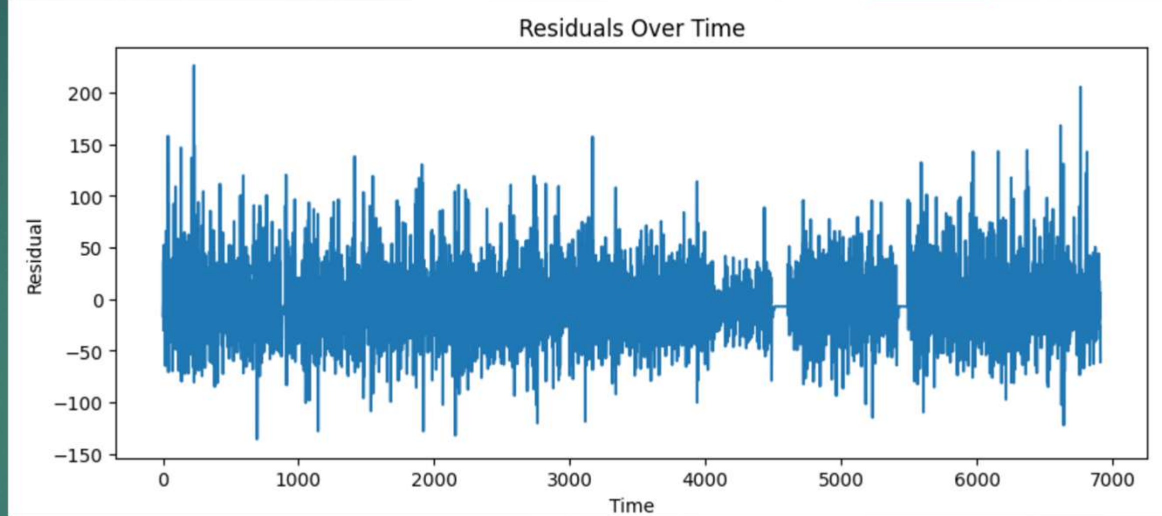
- Plot of Training and validation loss of BiLSTM model during training…

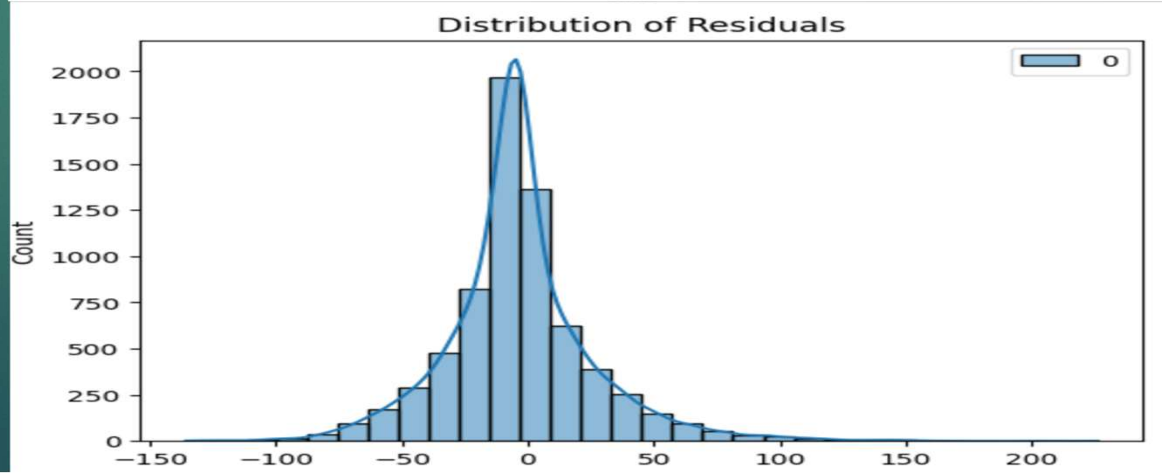




BiLSTM Model Diagnostics

- ❑ The residuals are scattered around 0 randomly, this strengthen the fact that residuals are independent.
- ❑ Also, distribution of residuals is very close to normal distribution.
- ❑ These two points indicates that residuals resembles white noise and model is reliable for forecasting.





A Team Project by:

- ❑ Jitesh Yadav (24N0066)
- ❑ Aditya Pandey (24N0064)
- ❑ Kunal Prajapat (24N0050)

Thank you !!