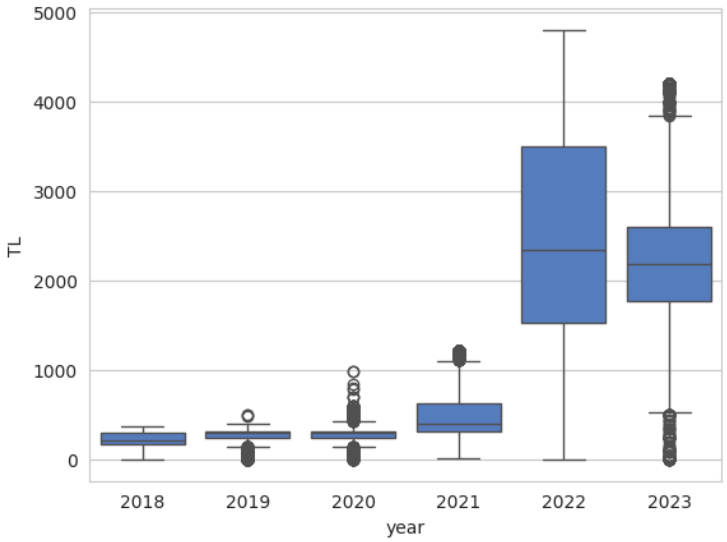
Ahmet Sayan
21903426

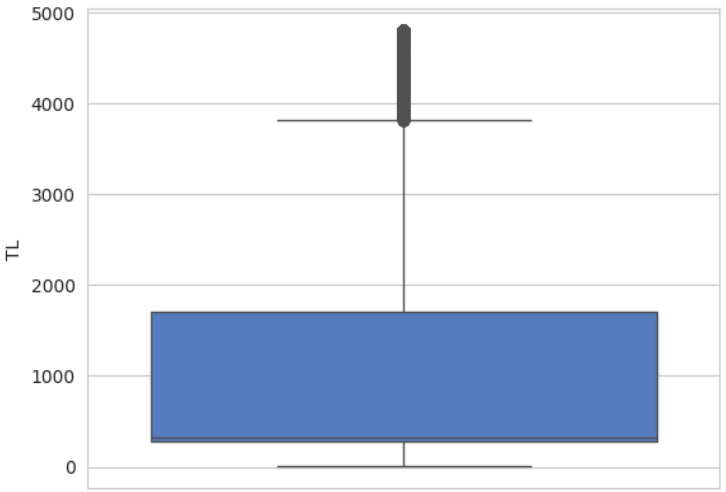
Man456 / Homework-1

1.Data Preparation

a)

Since the website doesn't allow downloading datasets for 2018-2023, I downloaded each year separately and concatenated them in the code. Then I searched for non-values and outliers. There were no non-values when I checked with the info function. For outliers, I checked boxplots for both yearly and overall since there is significant change over the years which seems like an outlier in the dataset but isn't. This kind of outliers might make some sense in the years we observed Corona.





I only eliminated points lower than 5 TL since these outliers make the MAPE value extremely large and meaningless. I didn't want to change other outliers much since deviation has been very large and insatiable over the years.

b)

By using feature engineering I evaluated columns and created day, year, hour and date time columns to evaluate later on.

c)

I created both covid_period and period columns for this part. First I made this part with only the covid_period dummy variable but then I realised that pre-corona and post-corona significantly differ from each other so I want to include this information as well.

This is the dataset I continued with after this part.

	Tarih	Saat	TL	EUR	Datetime	day	month	year	covid_period	period
0	2018-01-01	0	207.60	45.97	2018-01-01 00:00:00	Monday	1	2018	0	Pre-COVID
1	2018-01-01	1	205.34	45.47	2018-01-01 01:00:00	Monday	1	2018	0	Pre-COVID
2	2018-01-01	2	164.94	36.53	2018-01-01 02:00:00	Monday	1	2018	0	Pre-COVID
3	2018-01-01	3	154.52	34.22	2018-01-01 03:00:00	Monday	1	2018	0	Pre-COVID
4	2018-01-01	4	112.64	24.95	2018-01-01 04:00:00	Monday	1	2018	0	Pre-COVID
...	...	...	...	...	...	...	...	...	...	...
8755	2023-12-31	19	2499.67	76.74	2023-12-31 19:00:00	Sunday	12	2023	0	Post-COVID
8756	2023-12-31	20	2472.34	75.90	2023-12-31 20:00:00	Sunday	12	2023	0	Post-COVID
8757	2023-12-31	21	2472.33	75.90	2023-12-31 21:00:00	Sunday	12	2023	0	Post-COVID
8758	2023-12-31	22	1800.00	55.26	2023-12-31 22:00:00	Sunday	12	2023	0	Post-COVID
8759	2023-12-31	23	1345.15	41.30	2023-12-31 23:00:00	Sunday	12	2023	0	Post-COVID

52154 rows × 10 columns

2. Exploratory Data Analysis (EDA)

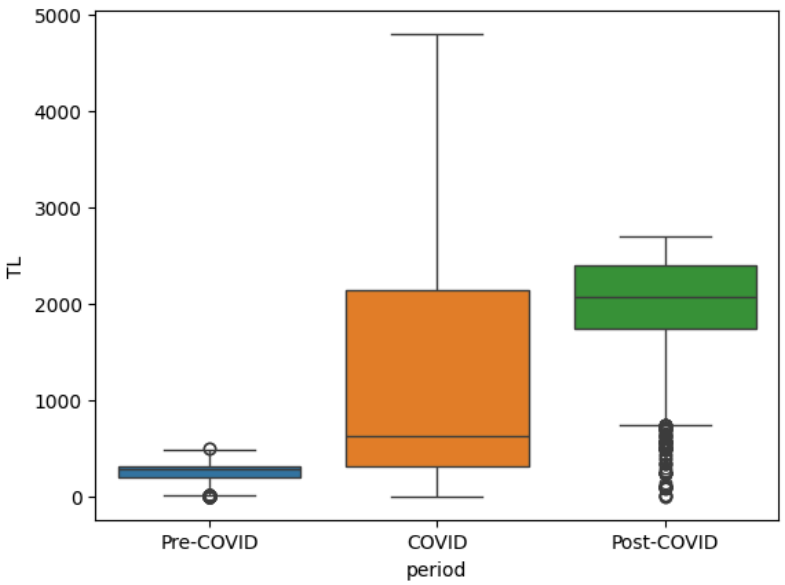
a)

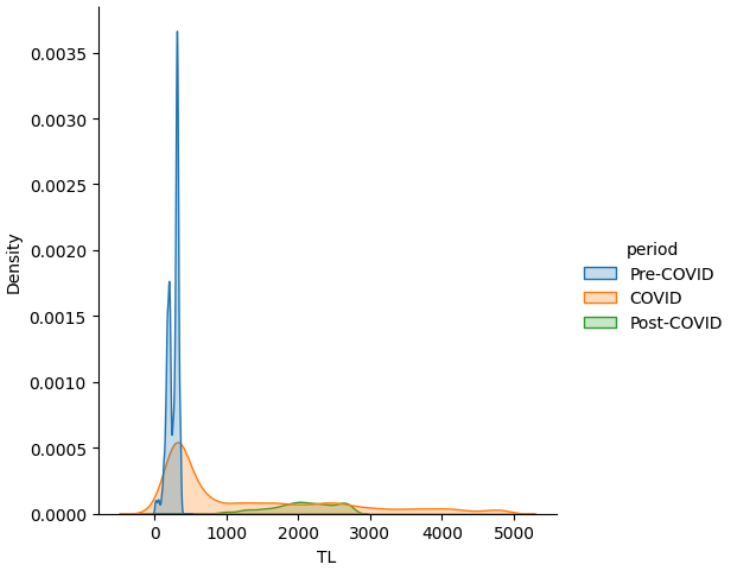
Statistics for electric prices according to the COVID period:

period	mean	median	std	min	max
Pre-COVID	254.125997	282.485	73.930555	5.01	500.0
COVID	1309.494557	636.000	1287.710260	5.01	4800.0
Post-COVID	2021.480575	2078.610	508.548267	10.00	2700.0

There are higher electric prices in the Covid and post-Covid periods compared to the pre-Covid period. Most importantly, there is a significant standard deviation in the Covid Period showing Covid caused high fluctuations.

b)



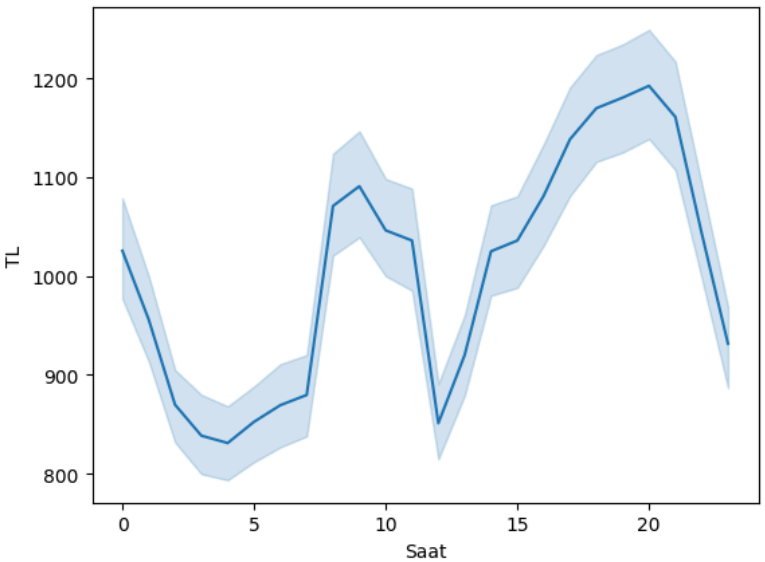


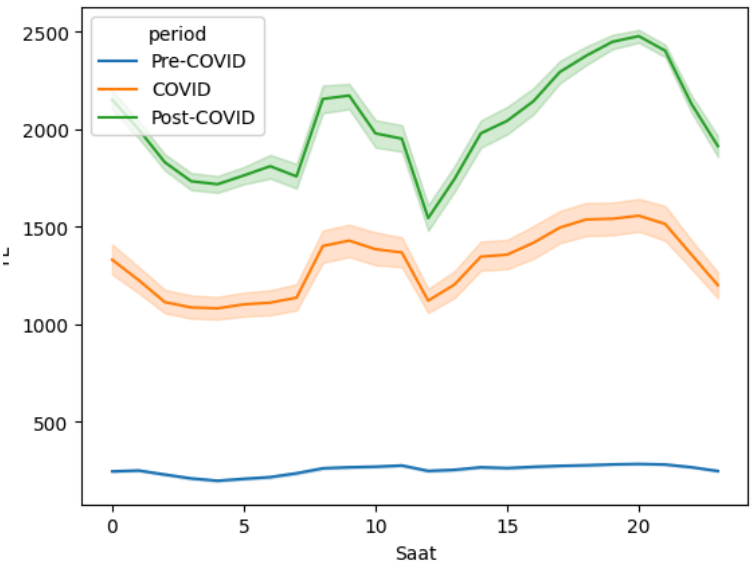
As I mentioned before, the COVID period has the highest standard deviation can be seen from the graphs. Both the histogram and boxplot of the Covid period are wider. I observed the mean value of price through periods increased.

c)

I mentioned that Corona's effect is significant over the electricity price so I wanted to check these graphs both with period information and not.

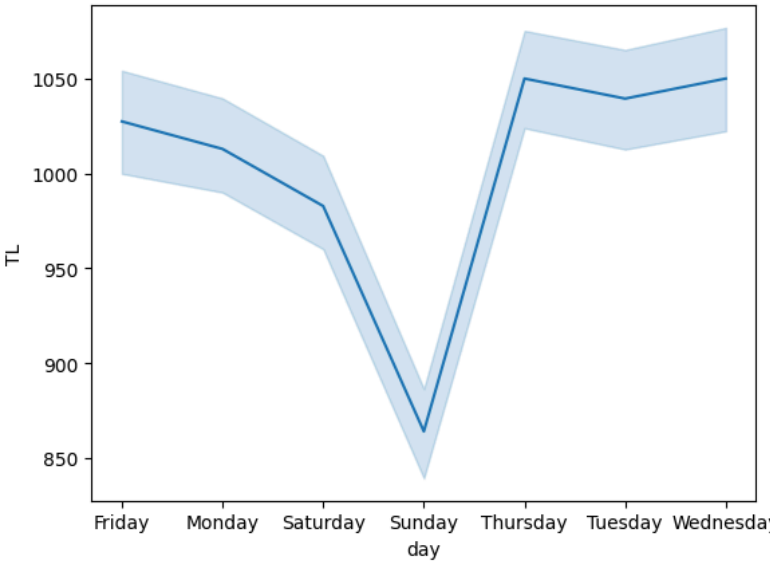
Hour-of-day effect

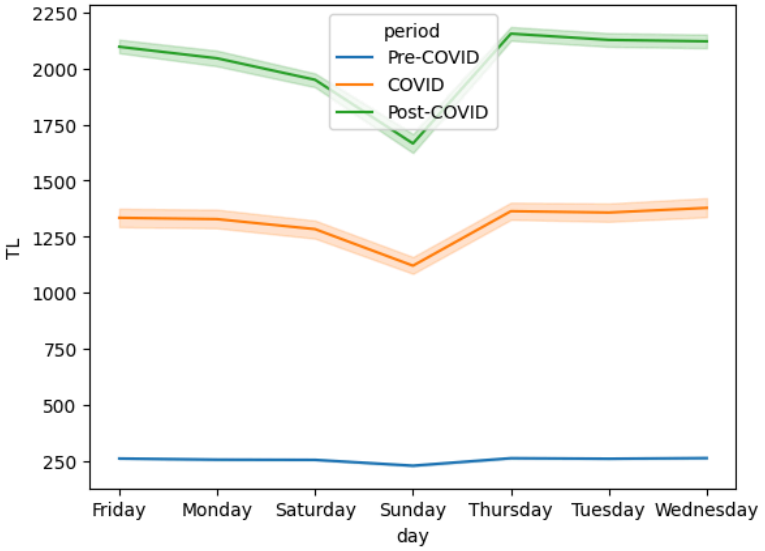




Even though hour information looks really important when I inspect all data (left graph), hour information doesn't seem that important for pre-COVID period (right graph). That gave me the idea to include the relation of these features in my model.

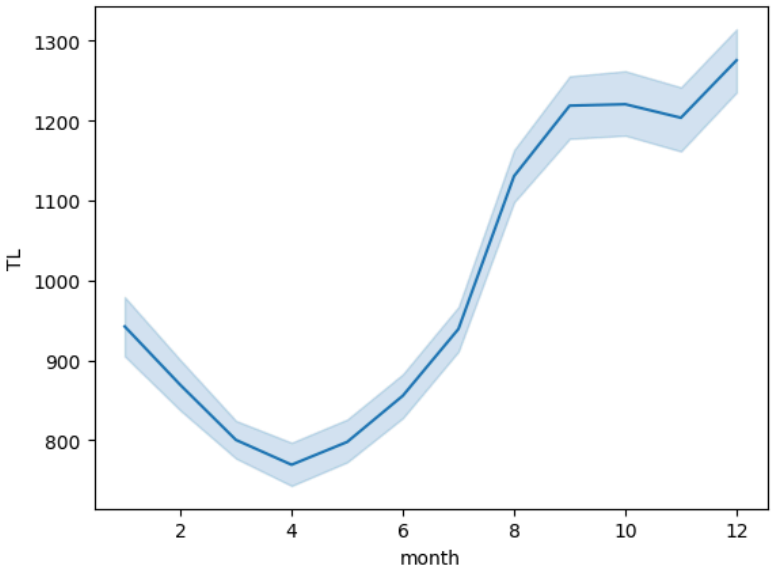
Day-of-week effect

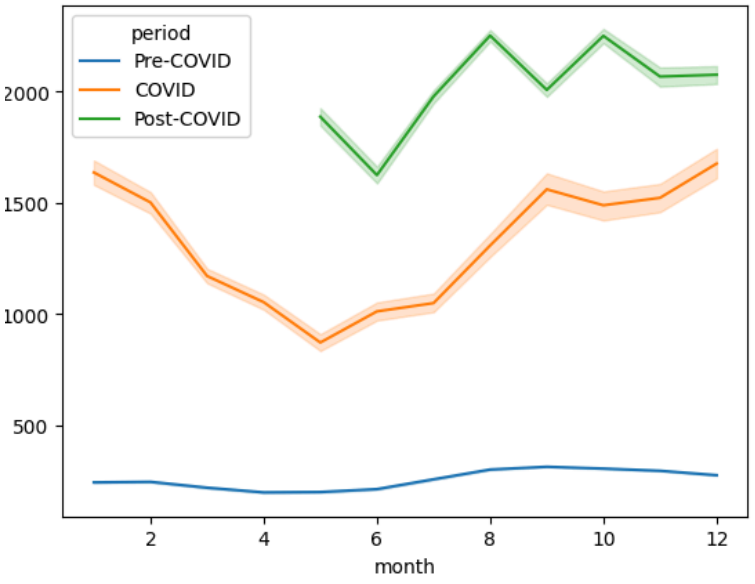




Sunday is significantly lower in terms of electricity price than other days of week. I also inspected it for different periods. Again, Pre-Covid doesn't have that effect on the price. It might also sense to have this feature relation in the model as well.

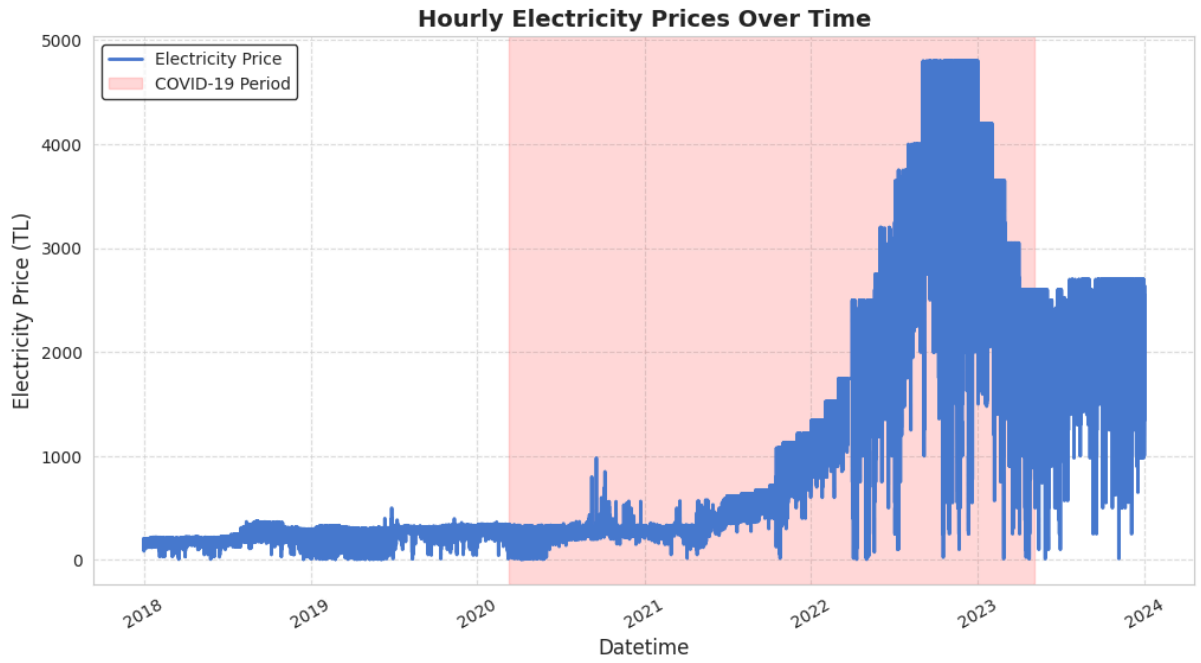
Monthly seasonality





There is the same pattern for monthly seasonality as well for Pre-COVID it doesn't have any effect. Additionally, Spring especially is cheaper than other months and fall is more expensive.

d)



As can be seen, Electricity Prices increased and started to fluctuate when Covid period started.

e)

I observed from all the graphs and statistics that electricity prices have increased and become very unstable during the Covid period. At the start, it didn't event change daily or monthly that much but after corona, even this informations started to affect the price.

3. Regression-Based Forecasting Model

a)

I divided my dataset into Test and Train.

b)

Date and time variables are categorical variables meaning that they should be turned into separate columns with one-hot encoding method. Therefore, first I applied One-hot encoding into my dataset. Then I fitted linear line but I wanted to inspect model for both Euro and TL to better see which one would look better.

OLS Regression Results

Dep. Variable:	TL	**R-squared:**	0.800
Model:	OLS	**Adj. R-squared:**	0.800
Method:	Least Squares	**F-statistic:**	3552.
Date:	Mon, 10 Mar 2025	**Prob (F-statistic):**	0.00
Time:	17:41:23	**Log-Likelihood:**	-3.1878e+05
No. Observations:	41723	**AIC:**	6.376e+05
Df Residuals:	41675	**BIC:**	6.381e+05
Df Model:	47		
Covariance Type:	nonrobust		

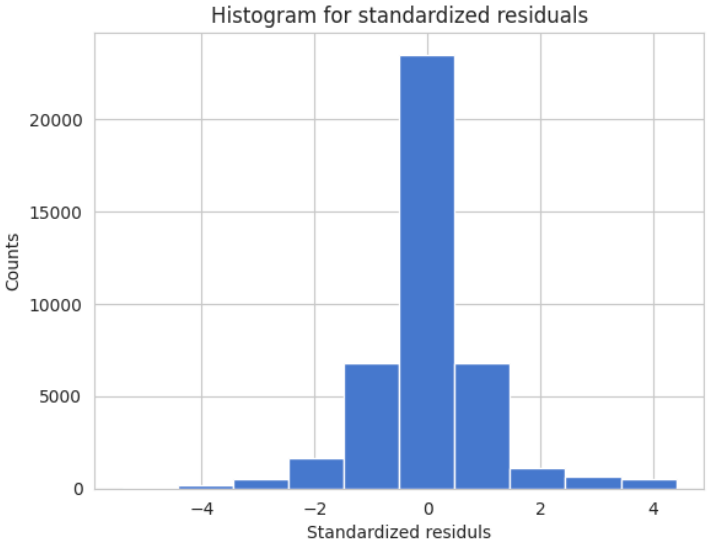
TL LinearModel Results

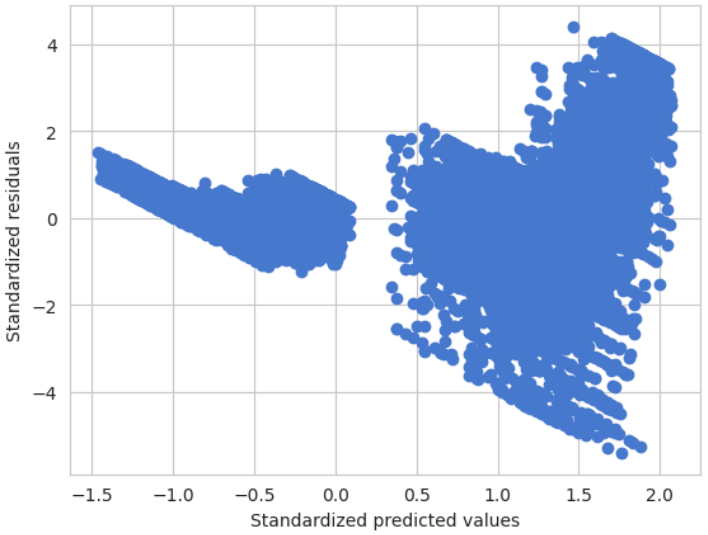
OLS Regression Results

Dep. Variable:	EUR	**R-squared:**	0.731
Model:	OLS	**Adj. R-squared:**	0.731
Method:	Least Squares	**F-statistic:**	2410.
Date:	Mon, 10 Mar 2025	**Prob (F-statistic):**	0.00
Time:	17:41:23	**Log-Likelihood:**	-1.9436e+05
No. Observations:	41723	**AIC:**	3.888e+05
Df Residuals:	41675	**BIC:**	3.892e+05
Df Model:	47		
Covariance Type:	nonrobust		

Euro LinearModel Results

R- Squared values for TL model is better so I continued with TL model. For this model I used year, hour, day, month and period informations but I would suggest better option later in this report.





I inspected assumptions of linear regression. Residuals looks normally distributed but there is funnel shape in the variance.

c)

These are the test results for TL model with Covid indicator

R-squared (R2): 0.807
Mean Absolute Percentage Error (MAPE): 0.872

These are the test results for TL model without Covid indicator

R-squared (R2): 0.786
Mean Absolute Percentage Error (MAPE): 0.763

I observed model with Covid indicator is better explanatory for variation in the dataset (R-squared) than model without Covid indicator. However, percentage error is decreased when I eliminated the covid indicator. I suggest that It is because of the multicollinearity between year feature and period feature. Since year feature already contains the information the period information within, it causes the problem in model.

d) Since I used categorical one hot encoding method and there were many categorical feature, it was hard to understand which coefficient is for which feature. But I can say that it increases the price and positive.

Better Model Suggestion

As I mentioned before adding relations of features might give better results so I tried this version where I added the relation and used these features: period-day, period-hour, period-month, year.

OLS Regression Results

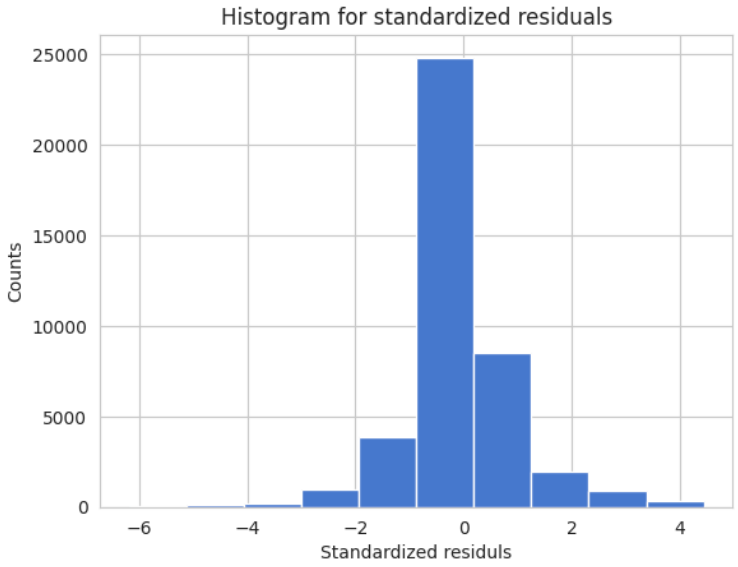
Dep. Variable:	TL	**R-squared:**	0.831
Model:	OLS	**Adj. R-squared:**	0.830
Method:	Least Squares	**F-statistic:**	1658.
Date:	Mon, 10 Mar 2025	**Prob (F-statistic):**	0.00
Time:	17:41:25	**Log-Likelihood:**	-3.1534e+05
No. Observations:	41723	**AIC:**	6.309e+05
Df Residuals:	41599	**BIC:**	6.320e+05
Df Model:	123		
Covariance Type:	nonrobust		

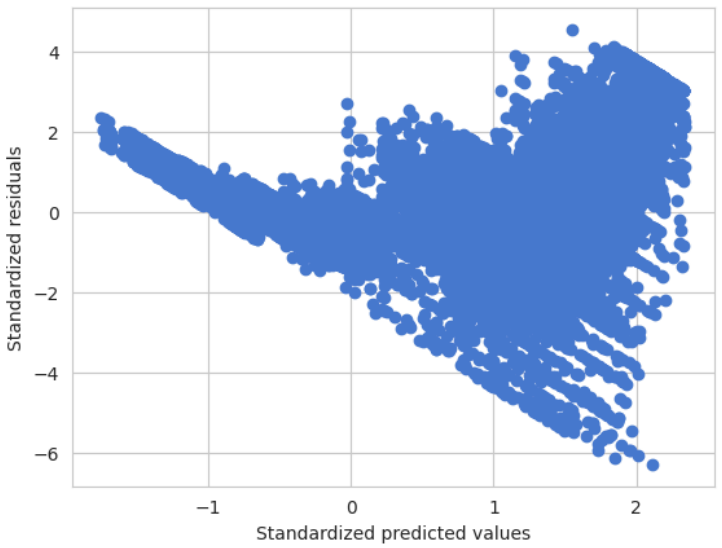
It gave slightly better R-squared for training set.

Also test set scores are better than previous model.

R-squared (R2): 0.837
Mean Absolute Percentage Error (MAPE): 0.760

I checked the graphs as well



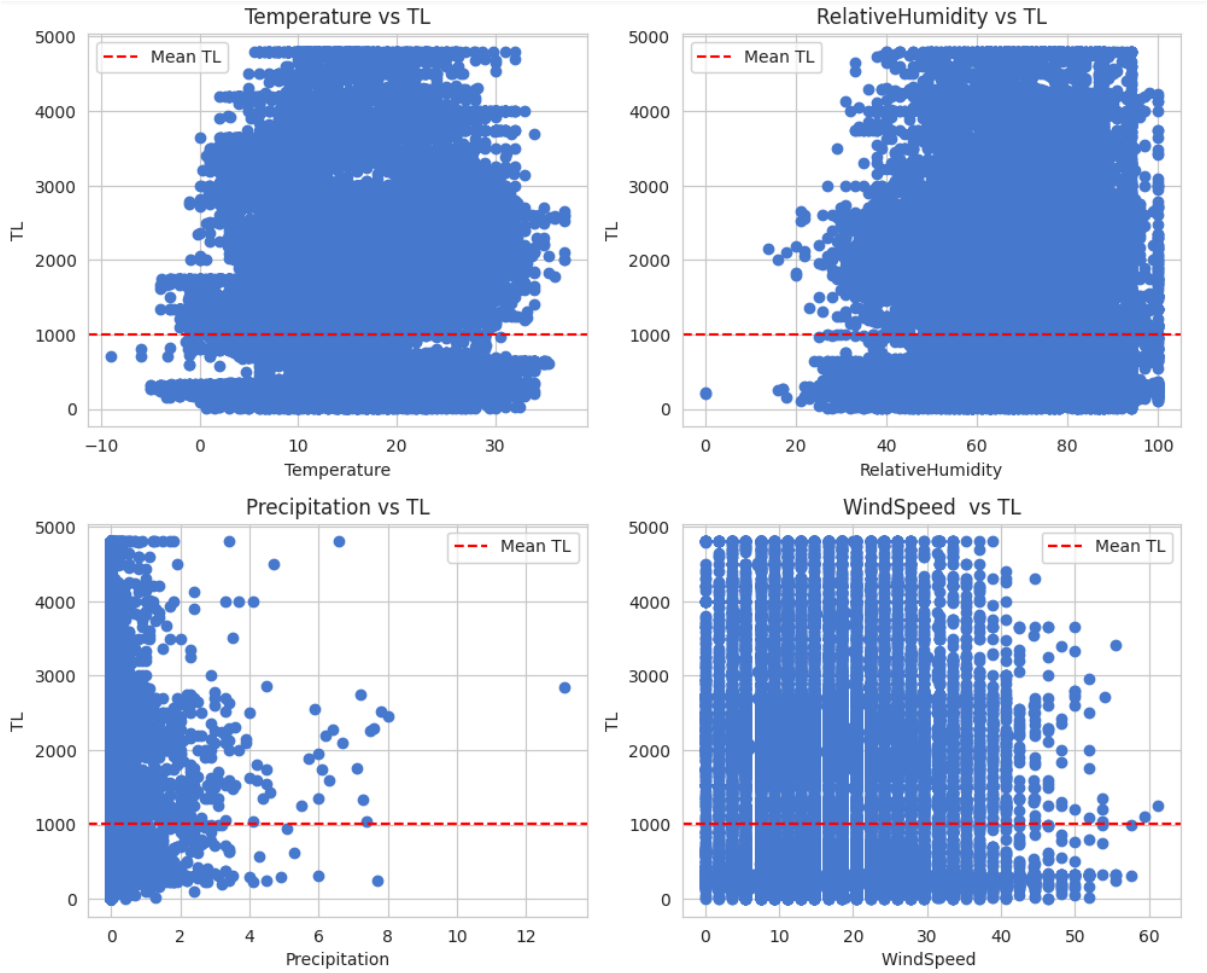


Residuals distributed normal but there is a funnel shape again in the variance. I didn't avoid the funnel shape.

Bonus Part

Weather Condition

For this part I important weather condition information from meteostat library and join with my dataset. I inspected hourly, temperature, precipitation, wind speed and humidity to see whether they have affect on electricity price or not.



From the graphs they doesn't seems significant features. Also adding them into my model doesn't increase R-squared much.

OLS Regression Results

Dep. Variable:	TL	**R-squared:**	0.832
Model:	OLS	**Adj. R-squared:**	0.832
Method:	Least Squares	**F-statistic:**	1623.
Date:	Mon, 10 Mar 2025	**Prob (F-statistic):**	0.00
Time:	17:42:01	**Log-Likelihood:**	-3.1515e+05
No. Observations:	41723	**AIC:**	6.306e+05
Df Residuals:	41595	**BIC:**	6.317e+05
Df Model:	127		

Also test set scores are not significantly better. Error rate even increased. P values of these features are significantly small except temperature but I think that because of the months already can keep the information of the weather.

R-squared (R2): 0.838
Mean Absolute Percentage Error (MAPE): 0.784