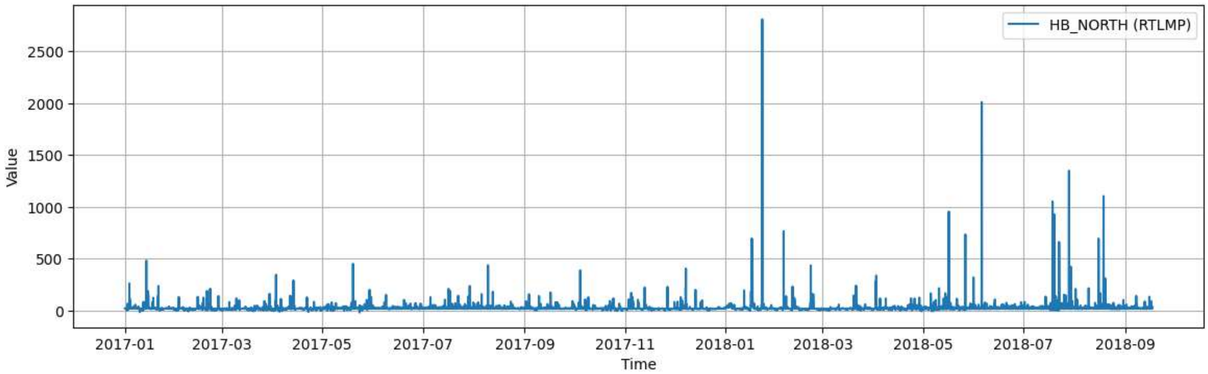
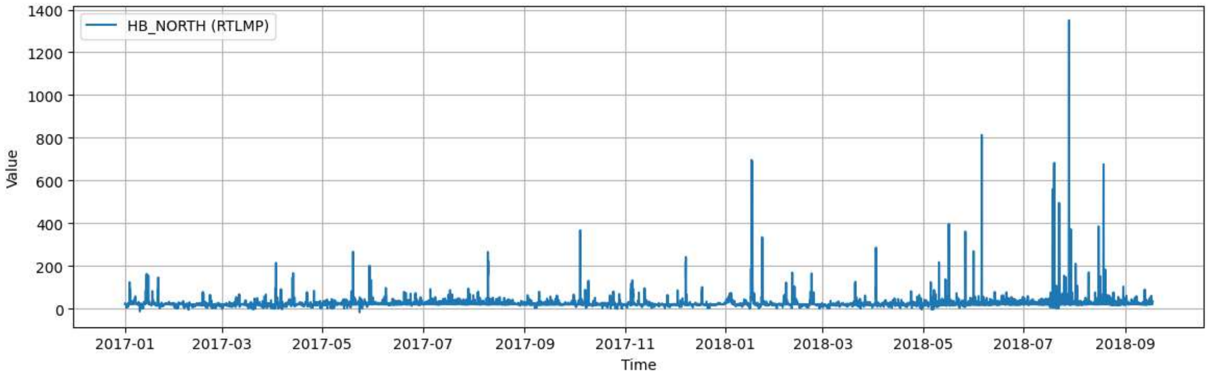
Assignment 3 - You Wu

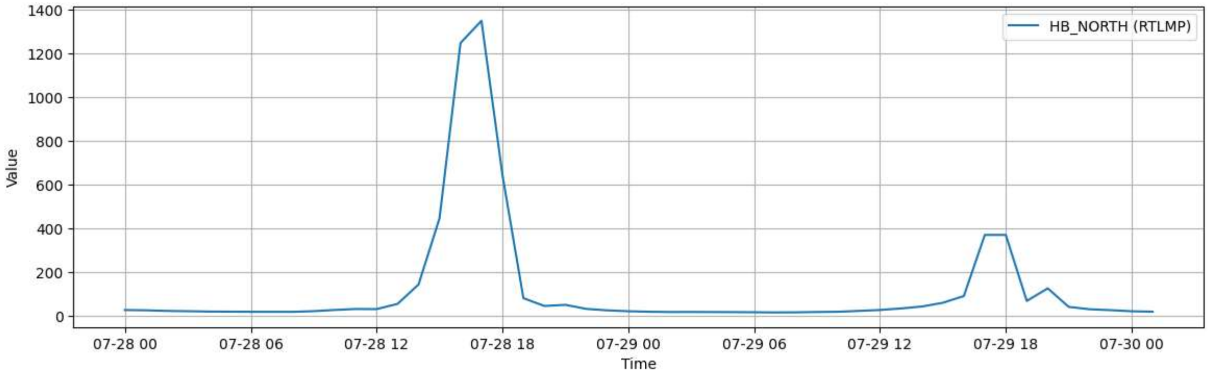
Before starting, check **data completeness.** There are two gaps and one repeat value in the timestamps, but that's because of the daylight-saving setting. There are only a few empty values. Because we are building forecast models, we use ffill instead of interpolation to fill the empty values.



There seem to be some outliers in the data. If a value deviates from its moving mean by more than 5 moving std, we consider it to be an outlier. We use a 30-hour hollow neighborhood to calculate the moving mean and moving std (15 hours before and 15 hours after, and that hour itself is not included). There're 169 outliers (1% of the total data) in RTLMP and no outliers in WIND_RTI, GENERATION_SOLAR_RT, and RTLOAD. We **drop the outliers and ffill them**. Below is the data without outliers.



There still seem to be some abnormal values. Here's an example.



We can see that though data on 07-28 and 07-29 is abnormal, it’s not an outlier. We should keep these abnormal values.

Split data into a train set and a test set.

Firstly, consider **OLS models**.

Use WIND_RTI, GENERATION_SOLAR_RT, and RTLOAD as number variables, and PEAKTYPE as a label variable. Use **the lag 1 term** of these variables to predict RTLMP.

GENERATION_SOLAR_RT and PEAKTYPE are not very useful in predicting RTLMP. OLS with WIND_RTI and RTLOAD has the lowest BIC. Lasso regression also suggests only using WIND_RTI and RTLOAD. The model is as below. ε is the residual.

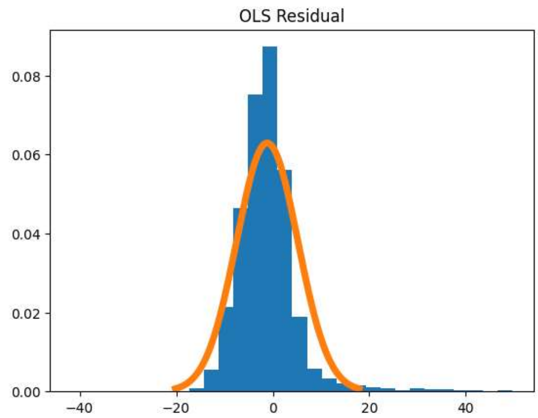
$$\text{RTLMP}_t = 3.0755 - 0.0009\ \text{WIND\_RTI}_{t-1} + 0.0006\ \text{RTLOAD}_{t-1} + \varepsilon_t$$

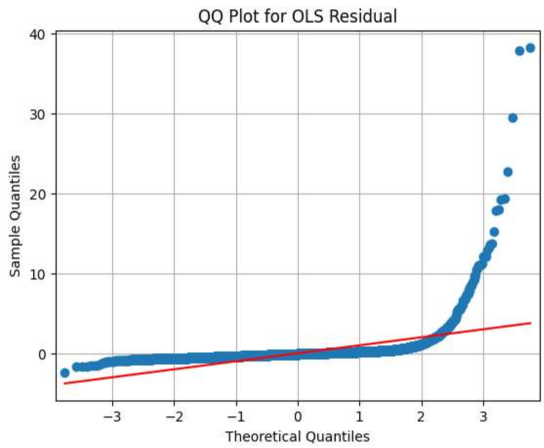
OLS Regression Results

Dep. Variable:	HB_NORTH (RTLMP)	R-squared:	0.149
Model:	OLS	Adj. R-squared:	0.149
Method:	Least Squares	F-statistic:	1049.
Date:	Sun, 21 May 2023	Prob (F-statistic):	0.00
Time:	15:12:11	Log-Likelihood:	-51133.
No. Observations:	11999	AIC:	1.023e+05
Df Residuals:	11996	BIC:	1.023e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.0755	0.878	3.501	0.000	1.354	4.797
ERCOT (WIND_RTI)	-0.0009	4e-05	-21.639	0.000	-0.001	-0.001
ERCOT (RTLOAD)	0.0006	1.86e-05	34.898	0.000	0.001	0.001

Omnibus:	25277.694	Durbin-Watson:	0.745
Prob(Omnibus):	0.000	Jarque-Bera (JB):	131901107.927
Skew:	18.212	Prob(JB):	0.00
Kurtosis:	515.345	Cond. No.	2.34e+05

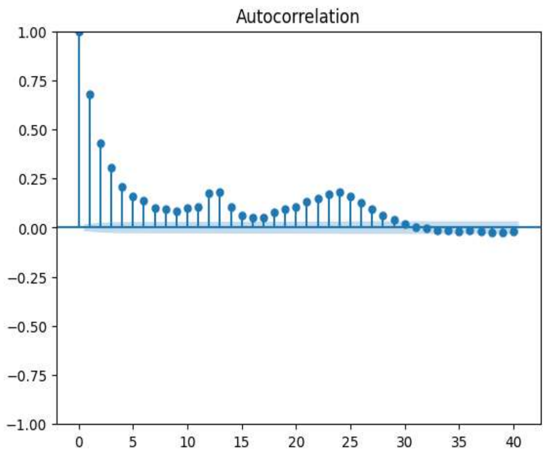


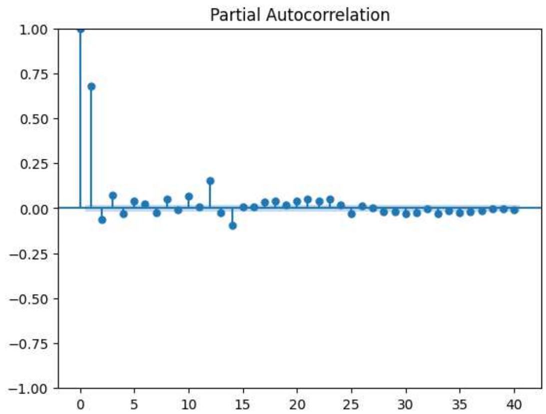


OLS residual is not really normal-distributed.

Secondly, consider **ARIMA models**.

Due to the ADF and KPSS tests, the price process is stationary.



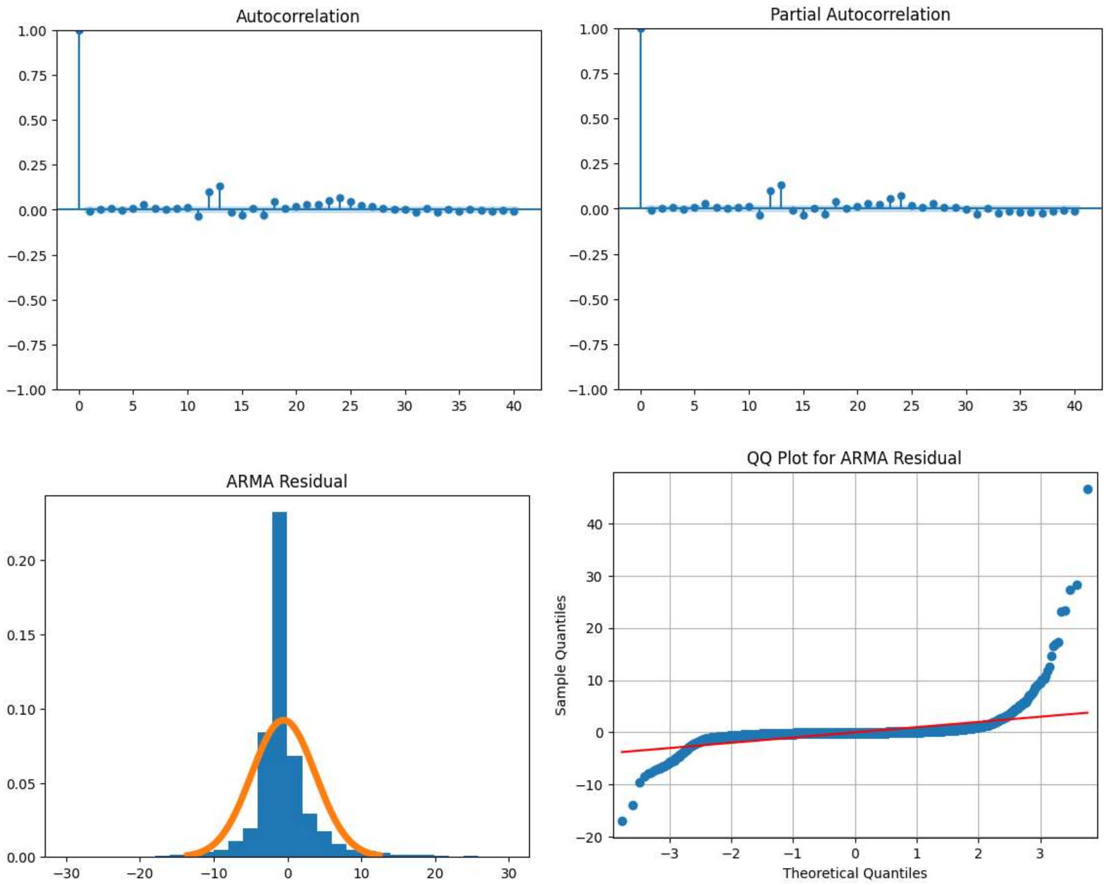


ACF and PACF of the price process show evidence of auto-correlation. Due to BIC, choose the ARMA(3,2) model. ε is the residual.

$$\mathrm{RTLMP}_t = 22.5850 - 0.9369\ \mathrm{RTLMP}_{t-1} + 0.3031\ \mathrm{RTLMP}_{t-2} + 0.4741\ \mathrm{RTLMP}_{t-3} + \varepsilon_t + 1.6871\varepsilon_{t-1} + 0.8196\varepsilon_{t-2}$$

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                      y   No. Observations:                11999
Model:                 ARIMA(3, 0, 2)   Log Likelihood              -48247.516
Date:                Sun, 21 May 2023   AIC                          96509.032
Time:                        15:14:20   BIC                          96560.780
Sample:                             0   HQIC                         96526.391
                              - 11999
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         22.5850      0.654     34.547      0.000      21.304      23.866
ar.L1         -0.9369      0.008   -114.344      0.000      -0.953      -0.921
ar.L2          0.3031      0.006     52.972      0.000       0.292       0.314
ar.L3          0.4741      0.008     63.185      0.000       0.459       0.489
ma.L1          1.6871      0.008    208.997      0.000       1.671       1.703
ma.L2          0.8196      0.009     88.993      0.000       0.802       0.838
sigma2       182.0500      0.329    553.354      0.000     181.405     182.695
===================================================================================
Ljung-Box (L1) (Q):                   1.26   Jarque-Bera (JB):          176794974.11
Prob(Q):                              0.26   Prob(JB):                          0.00
Heteroskedasticity (H):               2.91   Skew:                             16.22
Prob(H) (two-sided):                  0.00   Kurtosis:                        596.77
===================================================================================
```

Check the residuals of the ARMA model.



The Ljung-Box test is also rejected. There is no evidence of auto-correlation for ARMA residuals. The ARMA model has **residuals with lower values** than the OLS model. ARMA residuals are also more normal-distributed than OLS residuals. The ARMA model is effective.

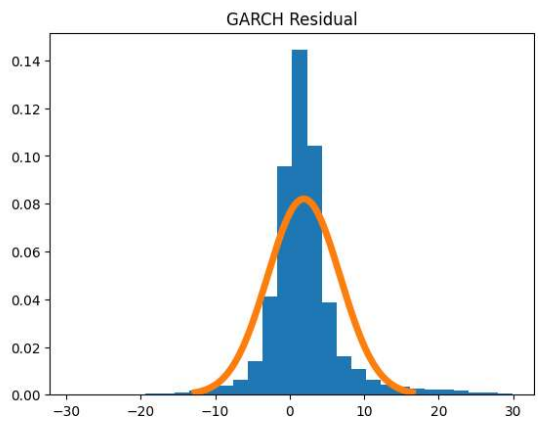
Also, consider the ARMAX(3,2) model with lag 1 term WIND_RTI and RTLOAD. It has a higher BIC.

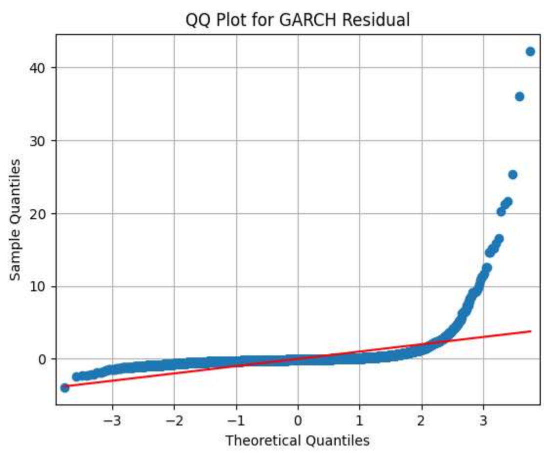
Thirdly, consider **GARCH models**.

Due to BIC, choose the ARX(3)-GARCH(2,0) model with lag 1 term RTLOAD. ε is the residual.

$$
\begin{aligned}
\text{RTLMP}_t &= 0.9873 + 0.2780\ \text{RTLMP}_{t-1} - 0.0416\ \text{RTLMP}_{t-2} - 0.0613\ \text{RTLMP}_{t-3} \\
&\quad + 0.00036\ \text{RTLOAD}_{t-1} + a_t \\
a_t &= \sigma_t \varepsilon_t \\
\sigma_t^2 &= 20.0141 + 0.9159 a_{t-1}^2 + 0.0841 a_{t-2}^2
\end{aligned}
$$

```
                          AR-X - ARCH Model Results
==============================================================================
Dep. Variable:      HB_NORTH (RTLMP)   R-squared:                       0.284
Mean Model:                     AR-X   Adj. R-squared:                  0.283
Vol Model:                      ARCH   Log-Likelihood:               -40632.9
Distribution:                 Normal   AIC:                           81281.8
Method:           Maximum Likelihood   BIC:                           81340.9
                                       No. Observations:                11996
Date:               Sun, May 21 2023   Df Residuals:                    11991
Time:                       15:15:38   Df Model:                            5
                                 Mean Model
================================================================================
                     coef    std err          t      P>|t|     95.0% Conf. Int.
--------------------------------------------------------------------------------
Const              0.9873      2.208      0.447      0.655   [ -3.340,  5.314]
HB_N...MP)[1]      0.2780      0.201      1.381      0.167   [ -0.117,  0.673]
HB_N...MP)[2]     -0.0416  4.054e-02     -1.027      0.304  [ -0.121,3.781e-02]
HB_N...MP)[3]     -0.0613  6.405e-02     -0.956      0.339  [ -0.187,6.428e-02]
ERCOT (RTLOAD) 3.5933e-04  1.212e-04      2.964  3.035e-03 [1.217e-04,5.969e-04]
                              Volatility Model
============================================================================
                 coef    std err          t      P>|t|       95.0% Conf. Int.
----------------------------------------------------------------------------
omega         20.0141     13.962      1.433      0.152   [ -7.351,  47.379]
alpha[1]       0.9159      0.306      2.990  2.788e-03   [  0.316,   1.516]
alpha[2]       0.0841  6.121e-02      1.373      0.170 [-3.592e-02,  0.204]
============================================================================
```



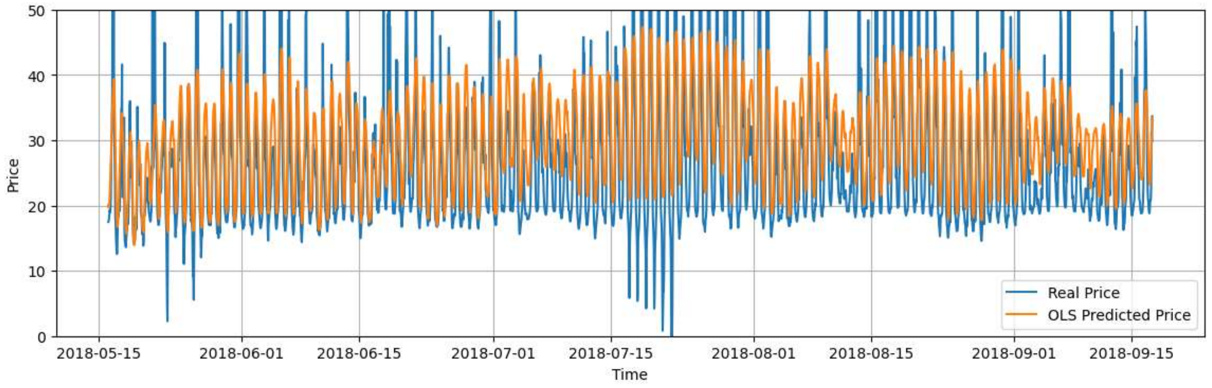


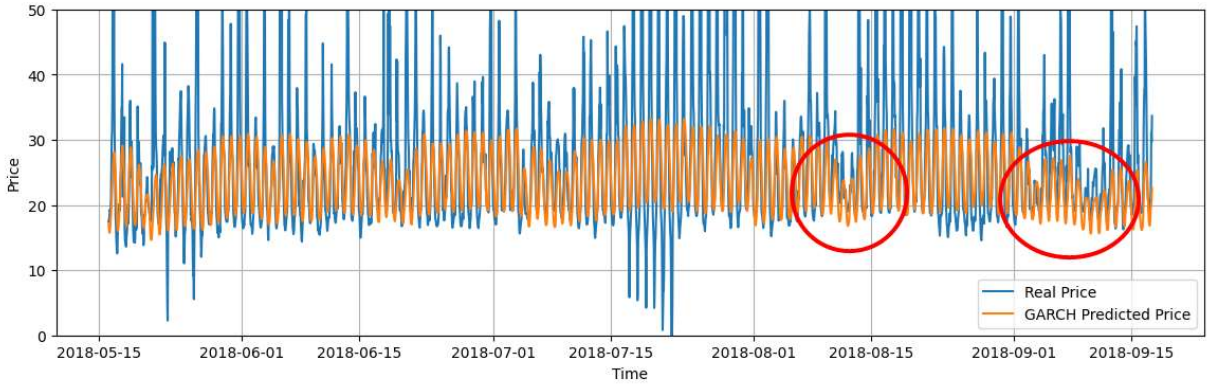
The GARCH model has **residuals with lower values** than the OLS model. GARCH residuals are also more normal-distributed than OLS residuals. The GARCH model is effective.

Lastly, **make predictions and calculate MSE** on the test set using the models.

```
OLS MSE:  3040.8380410248615
ARMA MSE:  3348.995360503033
GARCH MSE:  3213.4042339972075
```

The OLS model has the lowest MSE.





It is hard to predict the extreme values for both models. Generally, the OLS model performs better than the GARCH model. But in the periods where volatility changes, the GARCH model has a much better performance. Both models capture the pattern of the price and are useful to some extent.