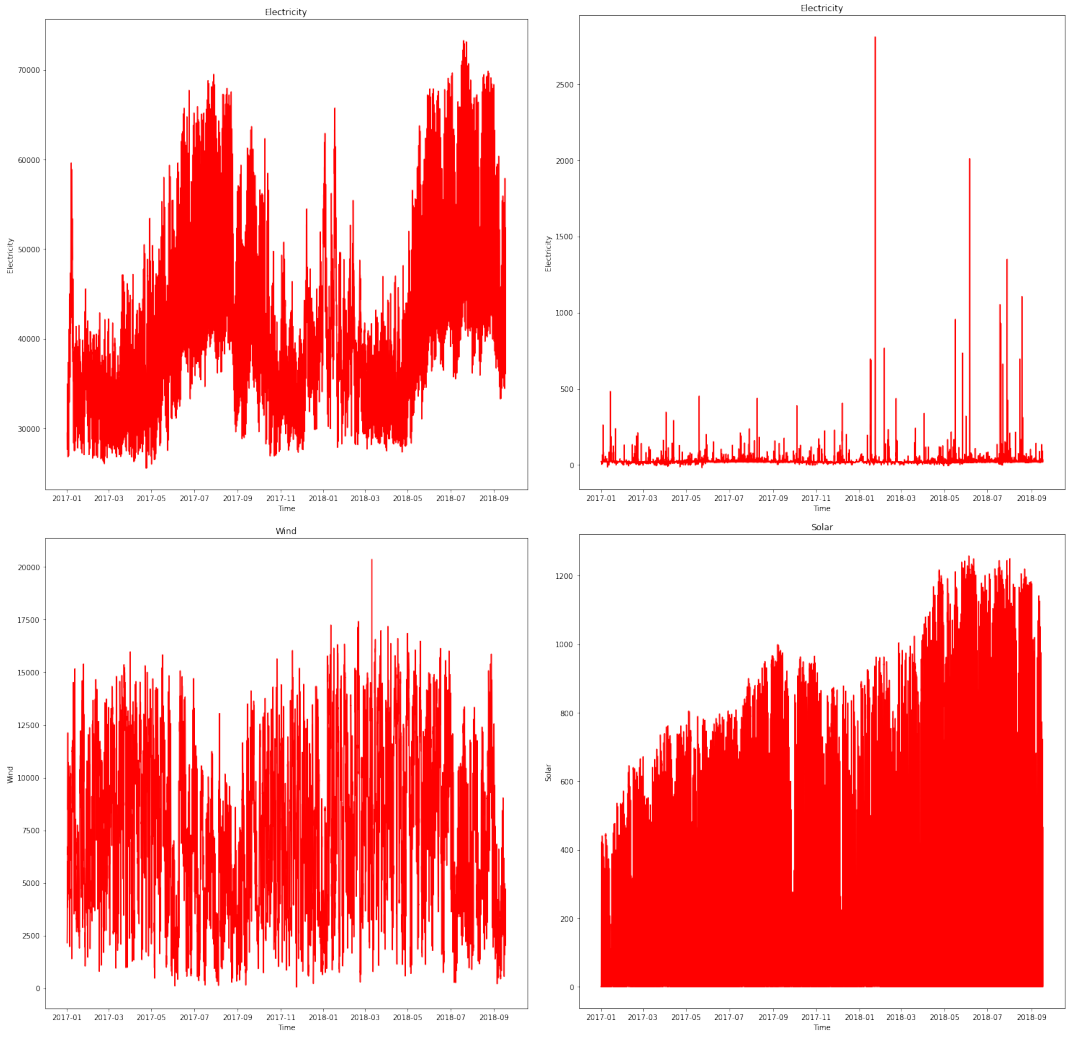
Guzman Homework Assignment Two

November 17, 2020

Assignment3 EDA and forecast model

1. **EDA**
   First, I look at the distributions of each of the columns by plotting the histograms. The real time LMP range from -18 to 2809. In the most of time, the LMP is less than 25. The distribution of RTLOAD looks like lognormal distribution. By log transformation the RTLOAD, the p value Shapiro test is very small. Thus, the distribution is not lognormal

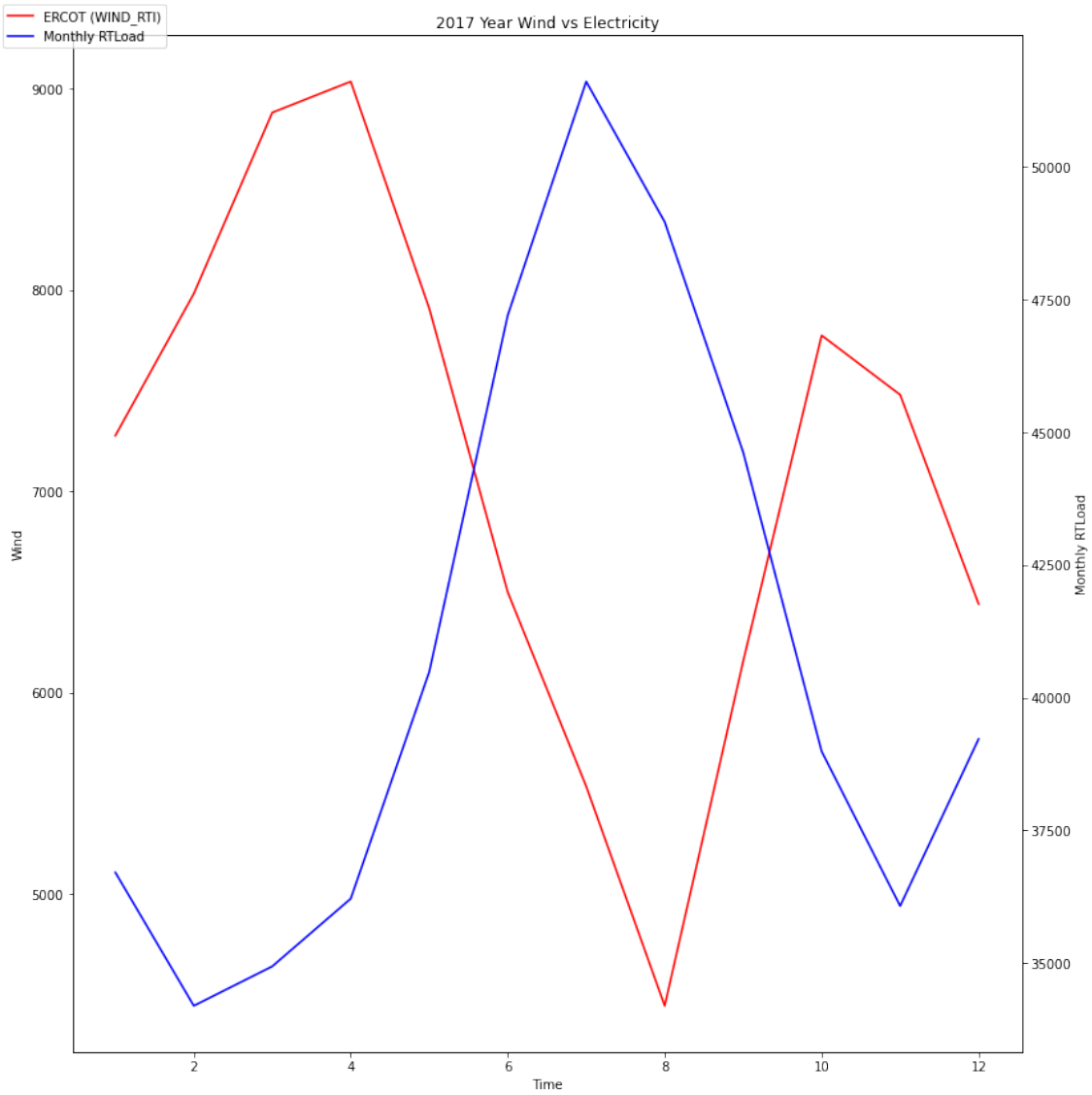
2. **Overall**



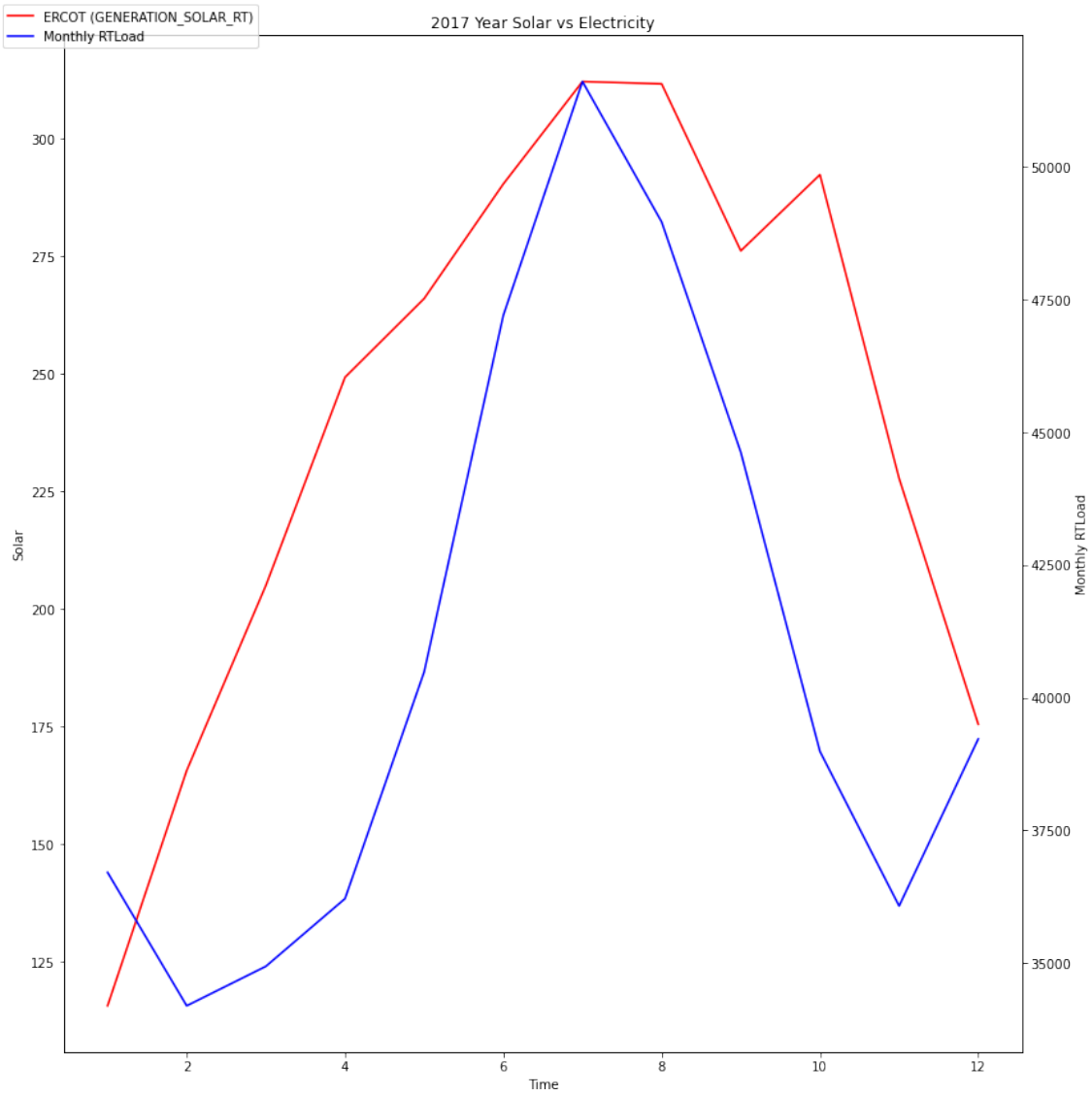
From the electricity load data, we can see the seasonality. The LMP in 2017 is stable, while LMP in

2018 has a lot of extreme and abnormal values. Wind generation data seems stationary. The solar generation has upward trend from 2017 to 2018.

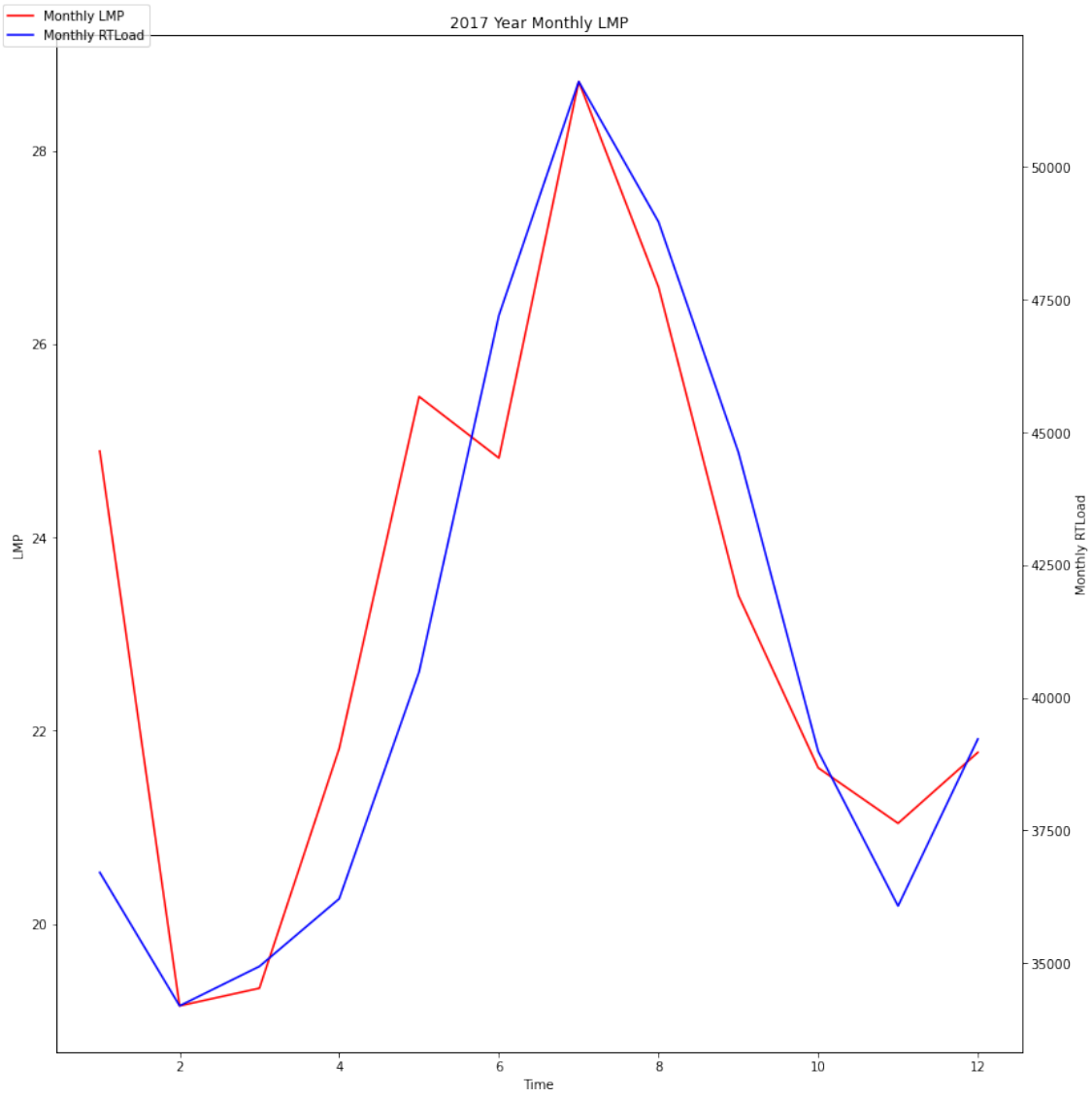
3. **2017 Monthly**



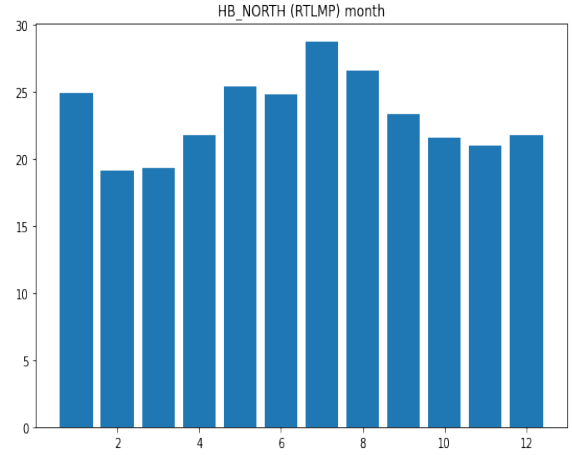
By plotting the monthly energy generation of electricity and wind in 2017, we witness strong inverse relationship between the wind generation and electricity generation. When wind generation increases, the electricity load decreases.When wind generation decreases, the electricity load increases. Electricity has the highest load in July, which may result from electricity consumption from AC in the summer. The wind generation peaks in April and reaches the minimum in August.

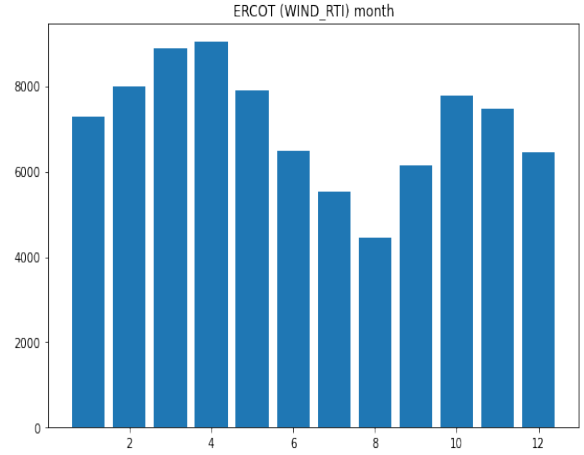


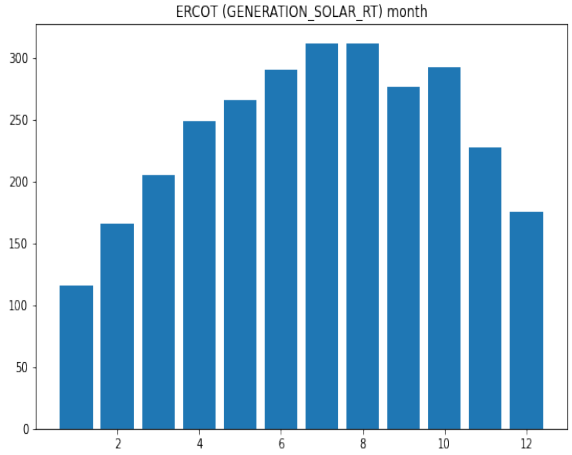
Solar generation may have positive correlation in the most of time except for the January and December. The solar generation reaches the highest point in July. It makes sense, because the sunlight in the summer amplifies the solar energy generation.

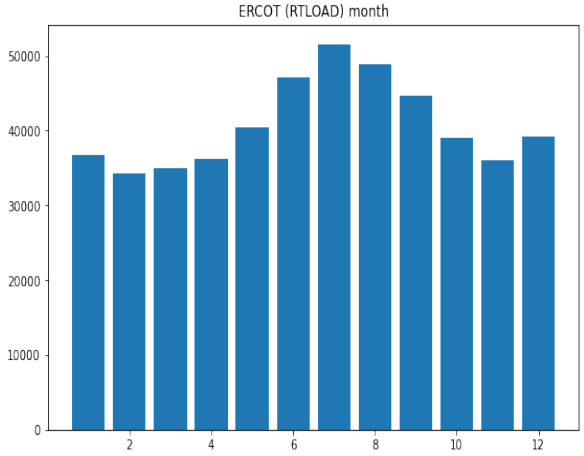


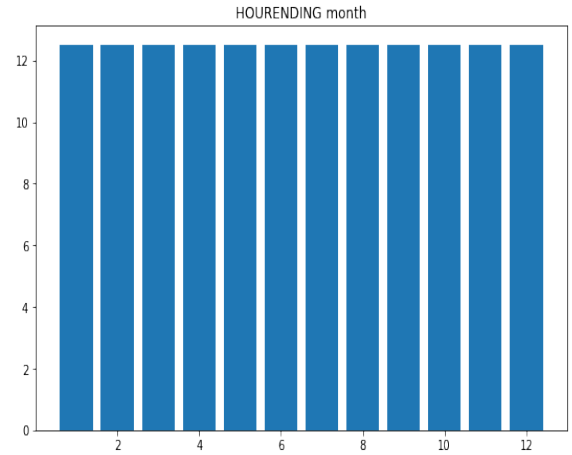
From the graph between RTLOAD and LMP, we can demonstrate that usually higher demand for electricity load may result in high electricity LMP. The electricity market is a demand-driven market

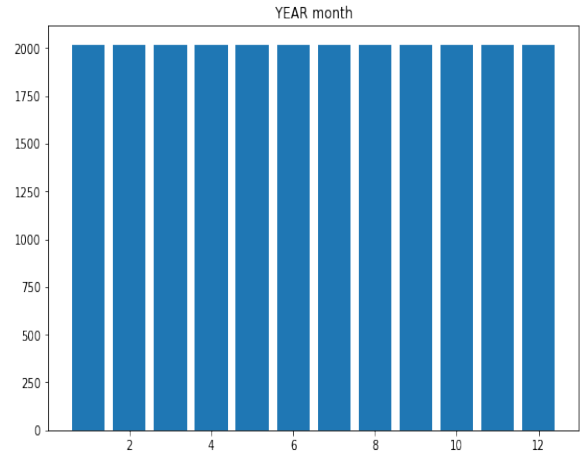




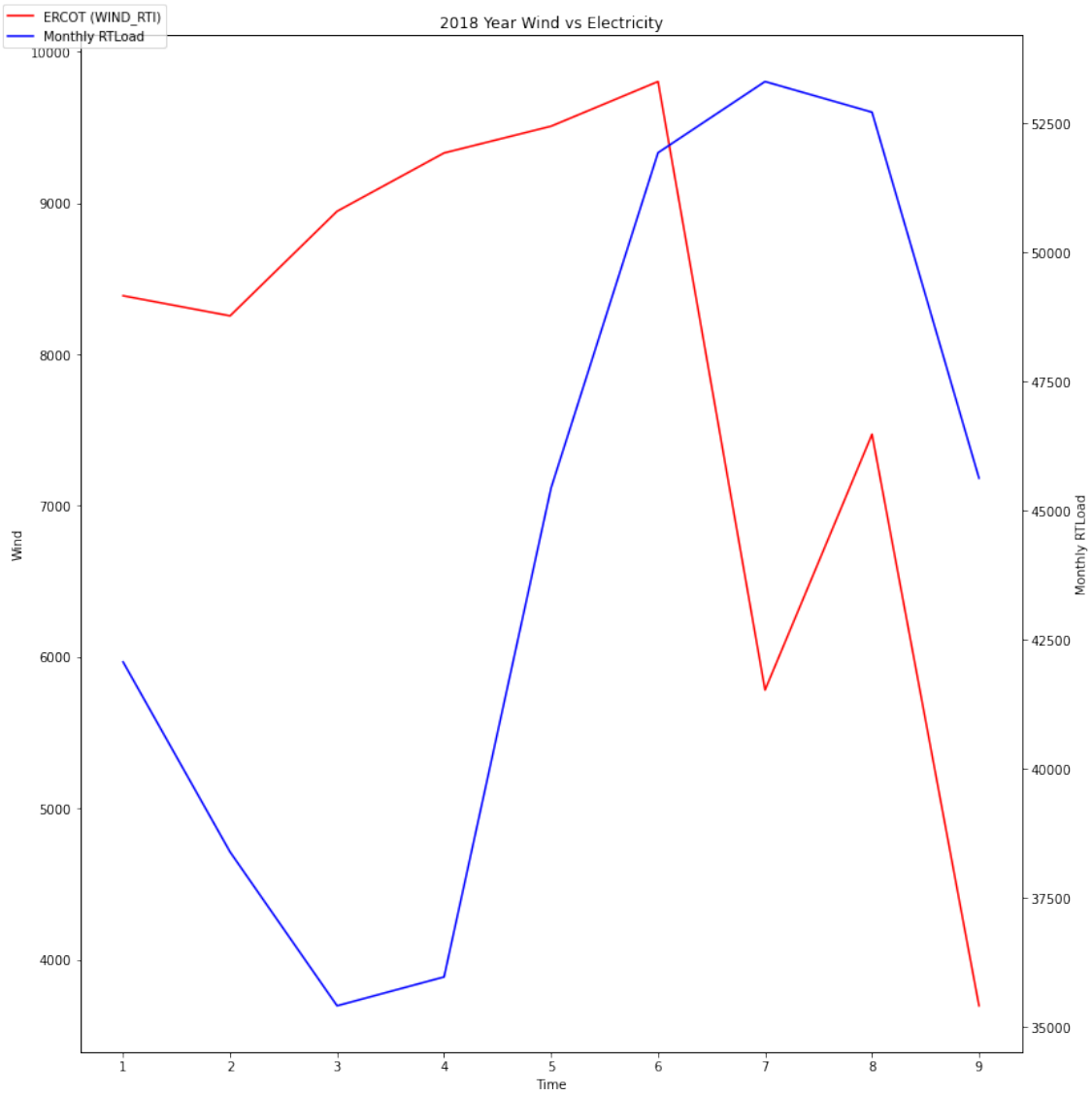




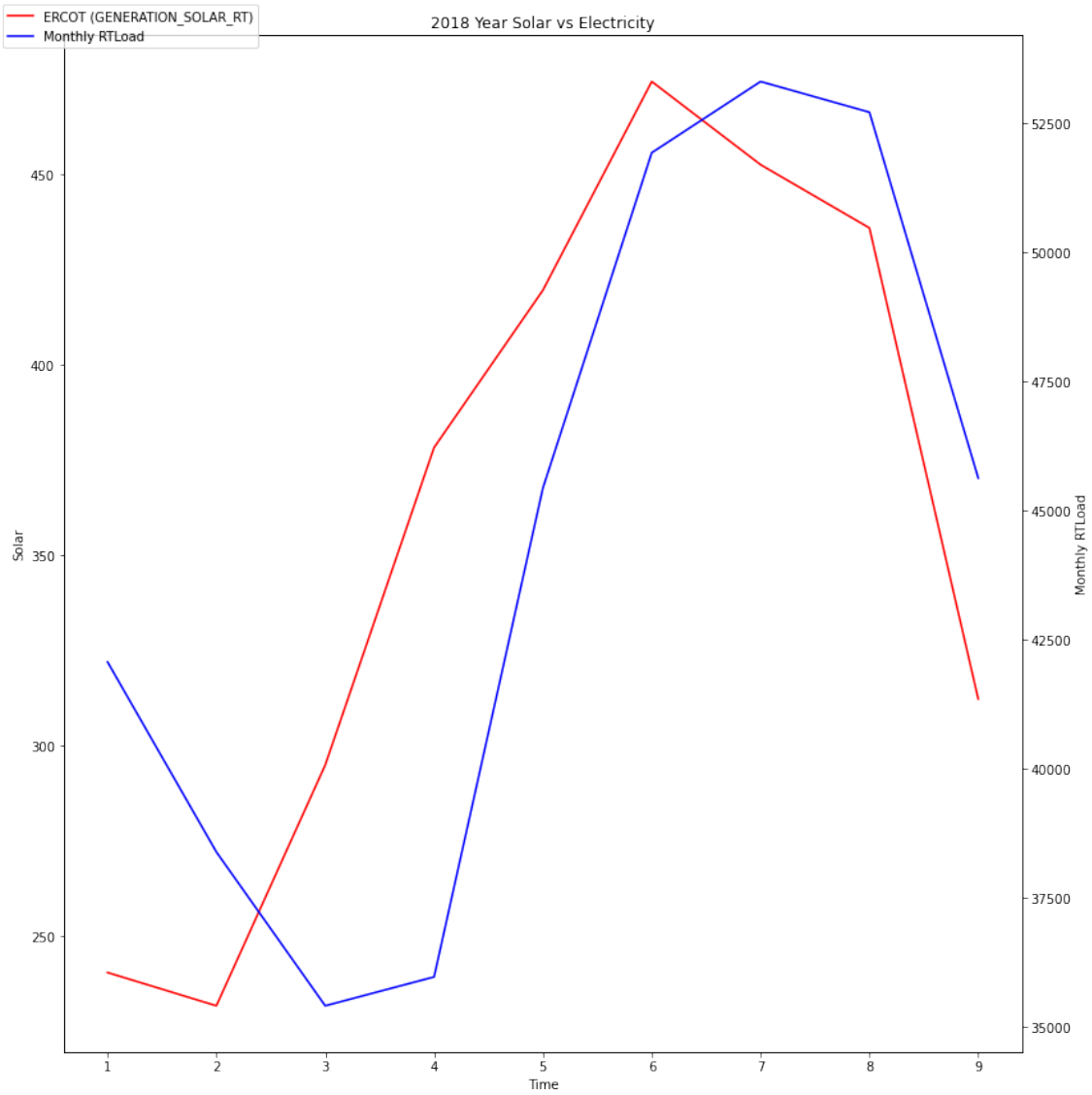




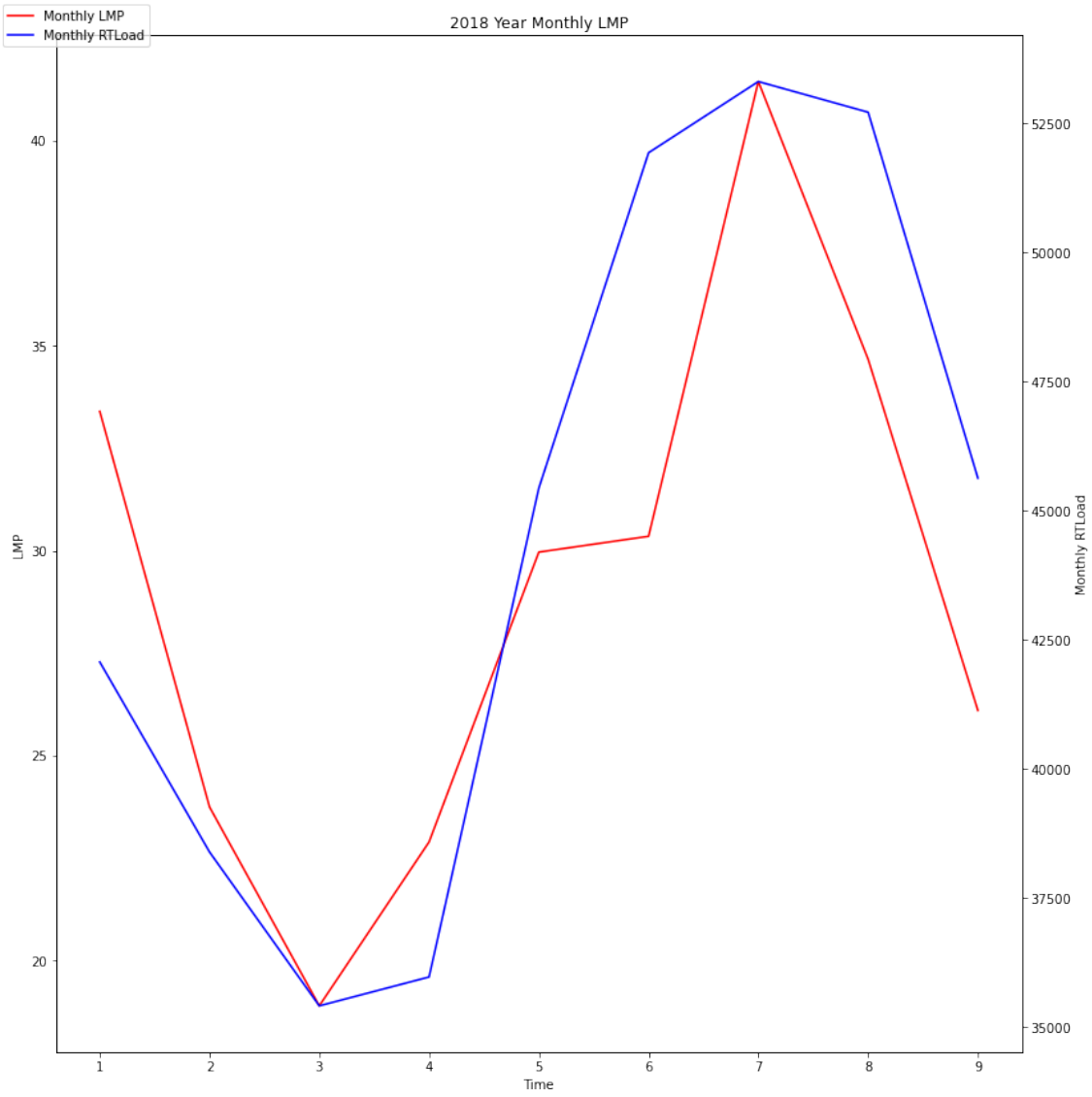
4. **2018 Monthly**



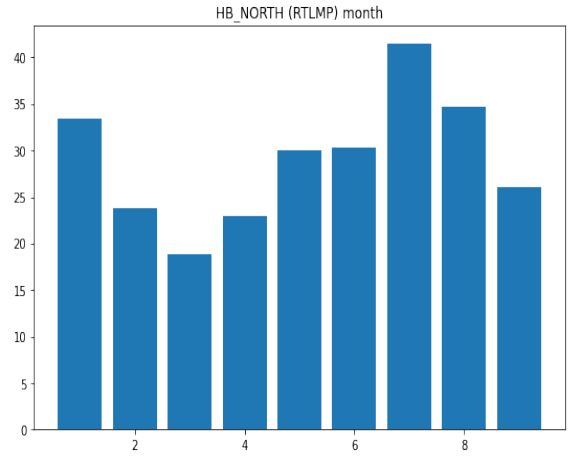
The 2018 data is not complete. By plotting the monthly energy generation of electricity and wind in 2018, we can't make any conclusion that strong inverse relationship between the wind generation and electricity generation anymore. The relationship between wind generation and electricity load is not totally inverse relationship just like 2017. They both decrease from Jan to Feb, increase from March to June, decrease from Aug to Sept. If we take a closer look at these two time series, we can precept that wind generation possibly lead the electricity load.

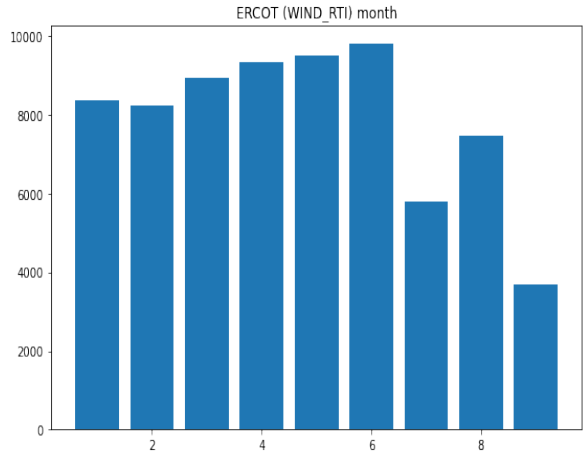


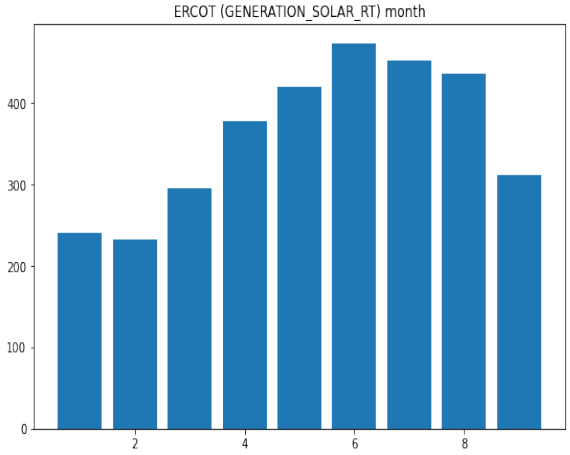
Solar generation may have positive correlation in the most of time just like 2017. If we take a closer look at these two time series, we can precept that solar generation possibly lead the electricity load.

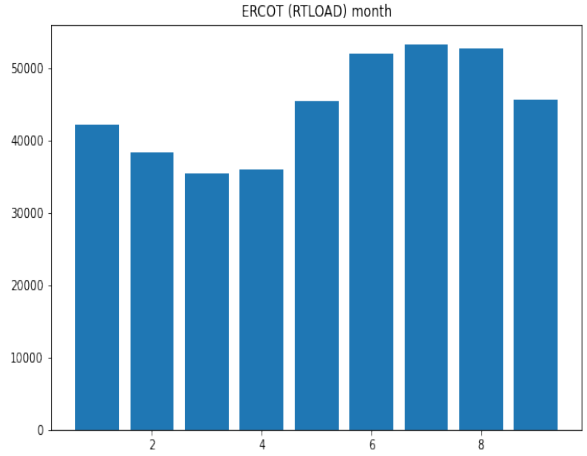


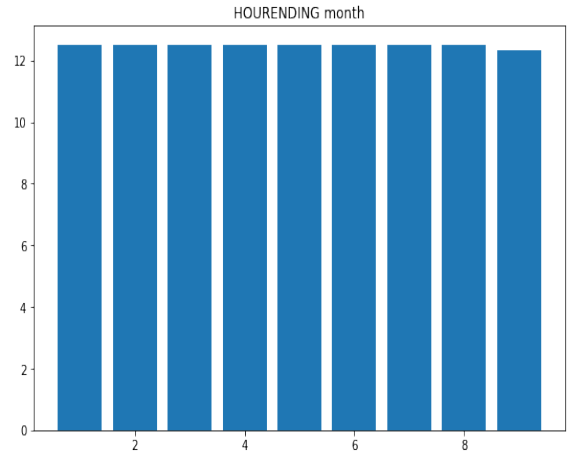
From the graph between RTLOAD and LMP in 2018, we can demonstrate that usually higher demand for electricity load may result in high electricity LMP just like 2017.

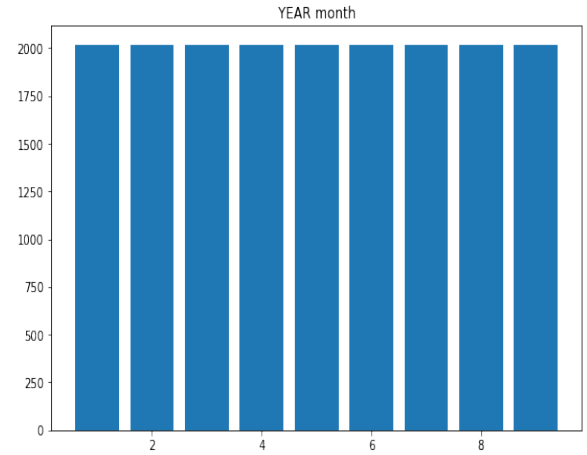

HB_NORTH (RTLMP) month


ERCOT (WIND_RTI) month

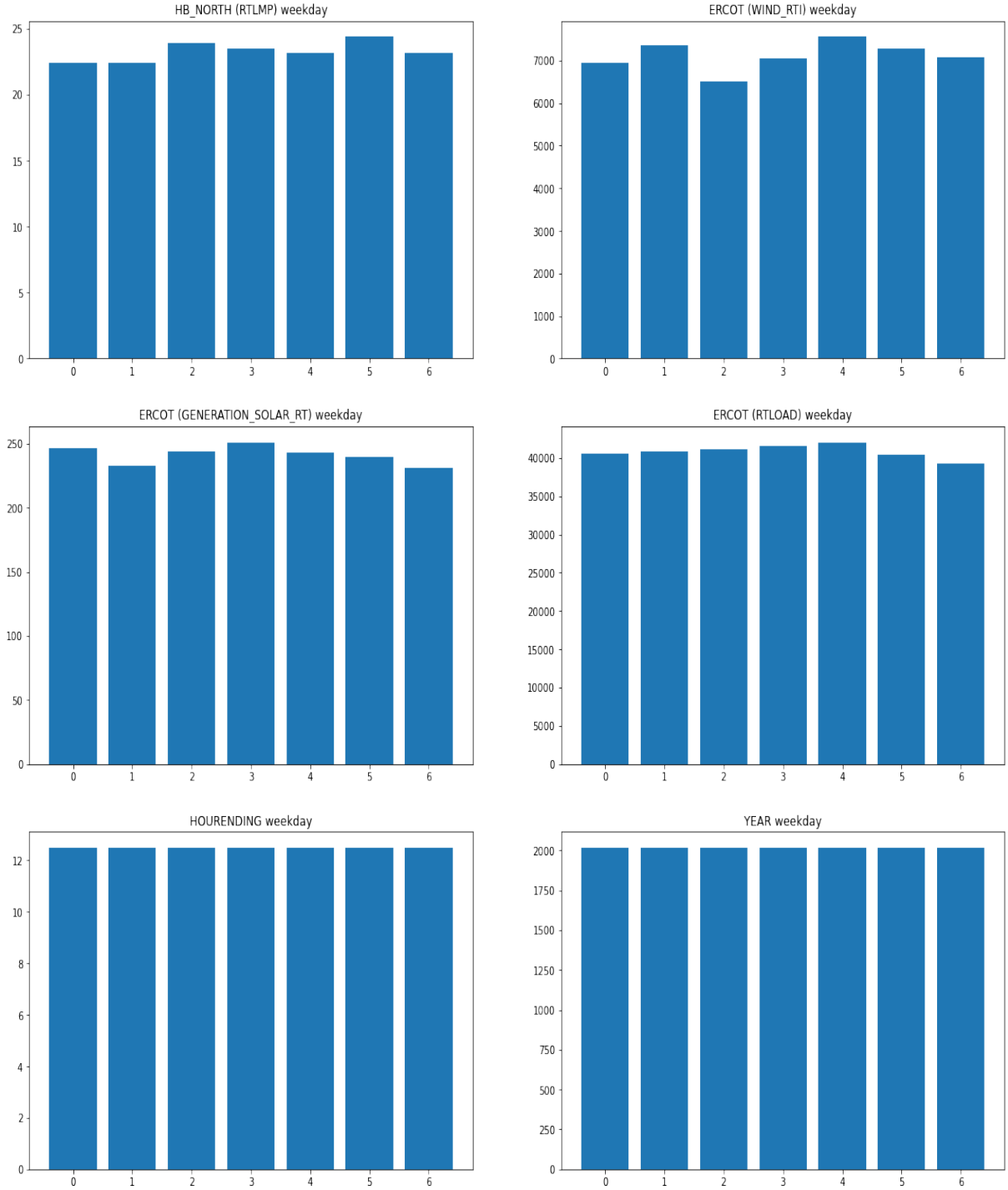

ERCOT (GENERATION_SOLAR_RT) month


ERCOT (RTLOAD) month


HOURENDING month

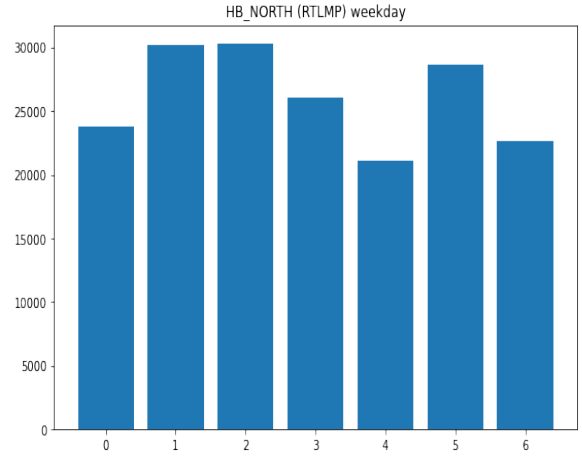

YEAR month

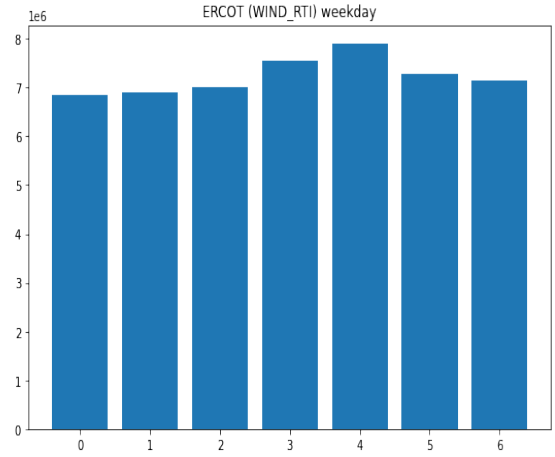
2017 week

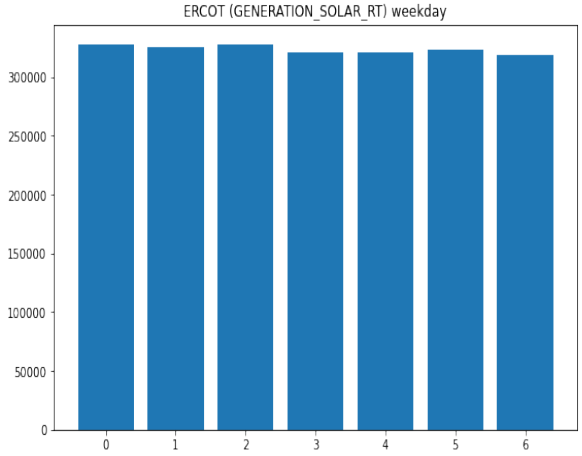


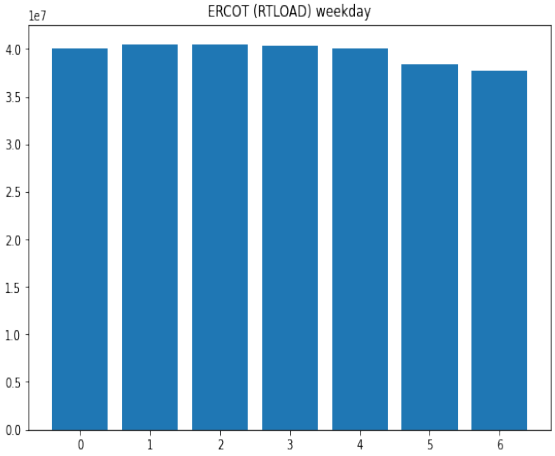
Saturday and Wednesday have the highest LMP. The solar generation and Electricity generation are relatively constant among these weekdays. Wind generation has the highest generation on Friday and lowest generation on Wednesday.

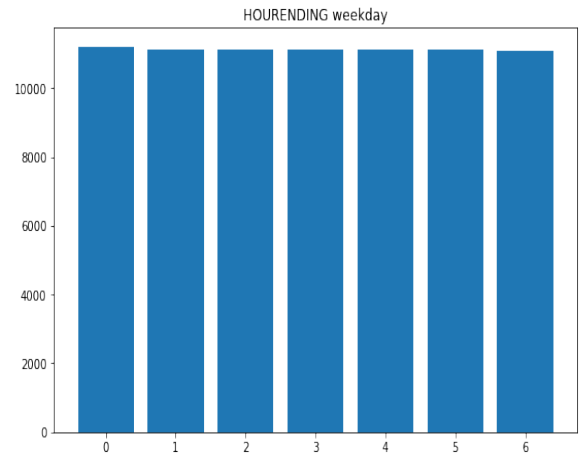
2018 week

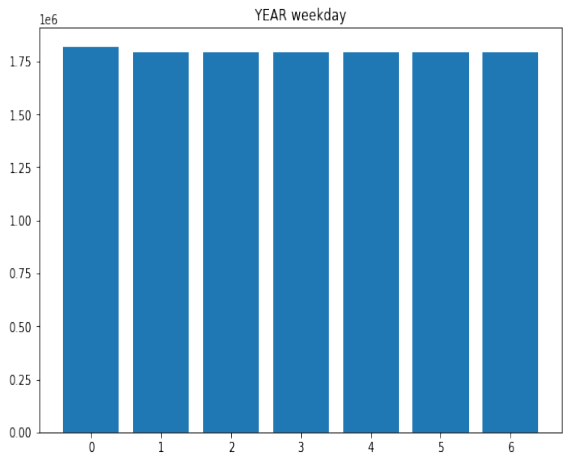






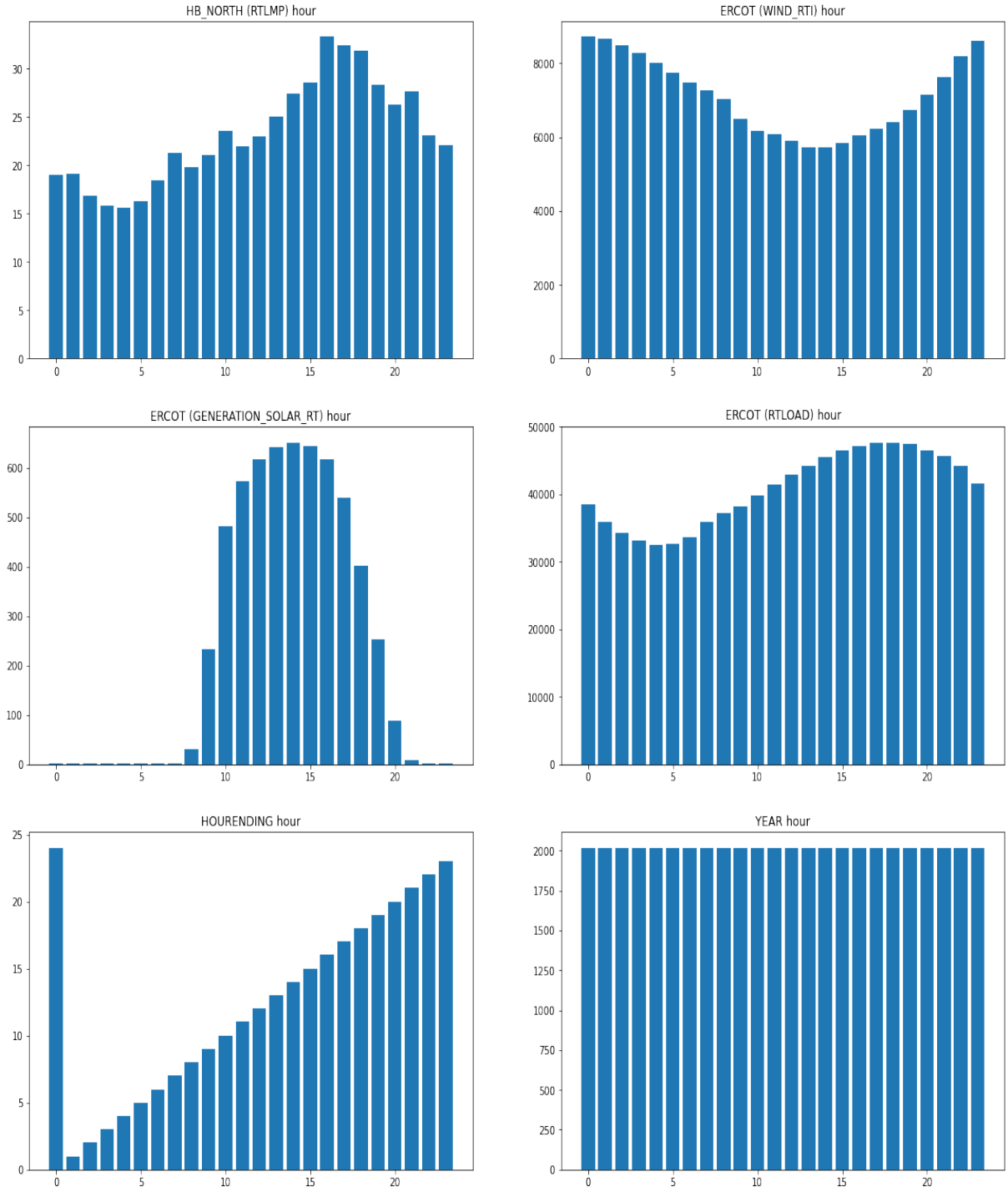






Tuesday, Wednesday and Saturday(like 2017) have the highest LMP, while Friday has the lowest LMP. The solar generation and Electricity generation are relatively constant among these weekdays. Wind generation has the highest generation on Friday.

2017 hour



solar generation only occurs between 8 AM to 8 PM. The wind generation has the local minimum at 2 PM. Electricity load peaks at 6PM and has the local minimum at 4 AM. The electricity consumption increases when people come back home after work. There are the least human activities at 4 AM when most of people are sleeping. The 2018 hour is the same as 2017 hour.

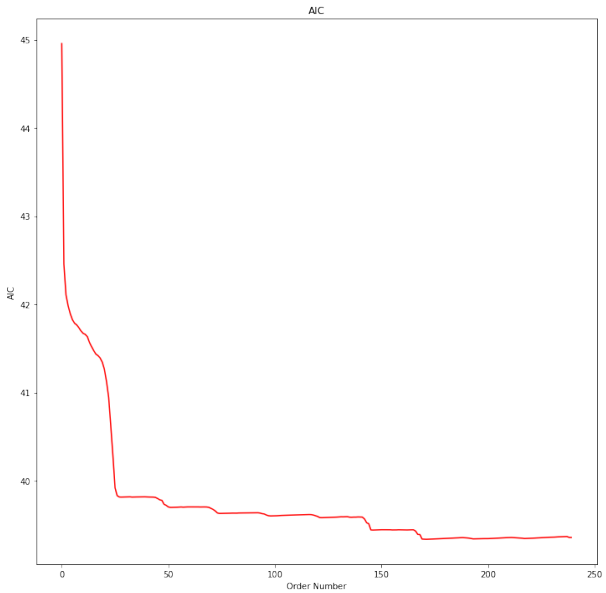
1. **forecast model: missing data transformation**
   The first step to building the forecast model is to cleaning data so that the data can be feed into the forecast model. When I check if the index is unique, I find that there are two 2017-11-05 02:00:00 in

the data frame. I decide to drop one of them. Besides, I check if there are some NaN in the data frame. There are five nan rows for ERCOT Wind generation. I decide to fill the NaN data with previous data in the time series, because I think the weather in hourly basis is constant. There may not be large change for wind speed in an hour. Moreover, there are four NaN rows in solar generation column. They are 2017-11-05 01:00:00 and 2017-11-05 02:00:00(duplicate). Based on what we learn from EDA, solar generation only happened between 8 AM to 8 PM. Thus, I fill the first three rows(1 AM and 2 AM) with zero. The final NaN row for solar generation is 2018-09-17 12:00:00. I fill the NaN with previous value, because the average solar generation for 12:00:00 is pretty close to the average solar generation for 11:00:00.
I transfer market date into days, since the month data and hour data are given in the data frame. Then, I transfer the peak type column into a dummy variable as well as year column and month column.

2. **forecast model: Vector Autoregression**
The first step for VAR is to run Granger Causality test that Check Granger Causality of all possible combinations of the time series. This test can if there is causality relationship between variables. For example, the past value of wind load at state t-1 may cause the RTLMP at state t. From the granger causation matrix, we can see that all variables cause the others, because all p values in the matrix are less than 0.05.This implies the Null Hypothesis that the coefficients of the corresponding past values is zero, that is, the X does not cause Y can be rejected.

	HB_NORTH (RTLMP)_x	ERCOT (WIND_RTI)_x	ERCOT (GENERATION_SOLAR_RT)_x	ERCOT (RTLOAD)_x
HB_NORTH (RTLMP)_y	1.0	0.0	0.0	0.0
ERCOT (WIND_RTI)_y	0.0	1.0	0.0	0.0
ERCOT (GENERATION_SOLAR_RT)_y	0.0	0.0	1.0	0.0
ERCOT (RTLOAD)_y	0.0	0.0	0.0	1.0

We run cointegration test for the data frame. Cointegration test is used to check if these time series are cointegrated, which means they have a long run, statistically significant relationship. It turns out all variables are cointegrated. After cointegration test, the ADF test is necessary to check the stationary of the time series. Luckily, all time series are stationary. The next step is to determine the order of VAR. By trying different order number, I pick the order with the lowest AIC. I decide to use 172 as order number for the VAR model.

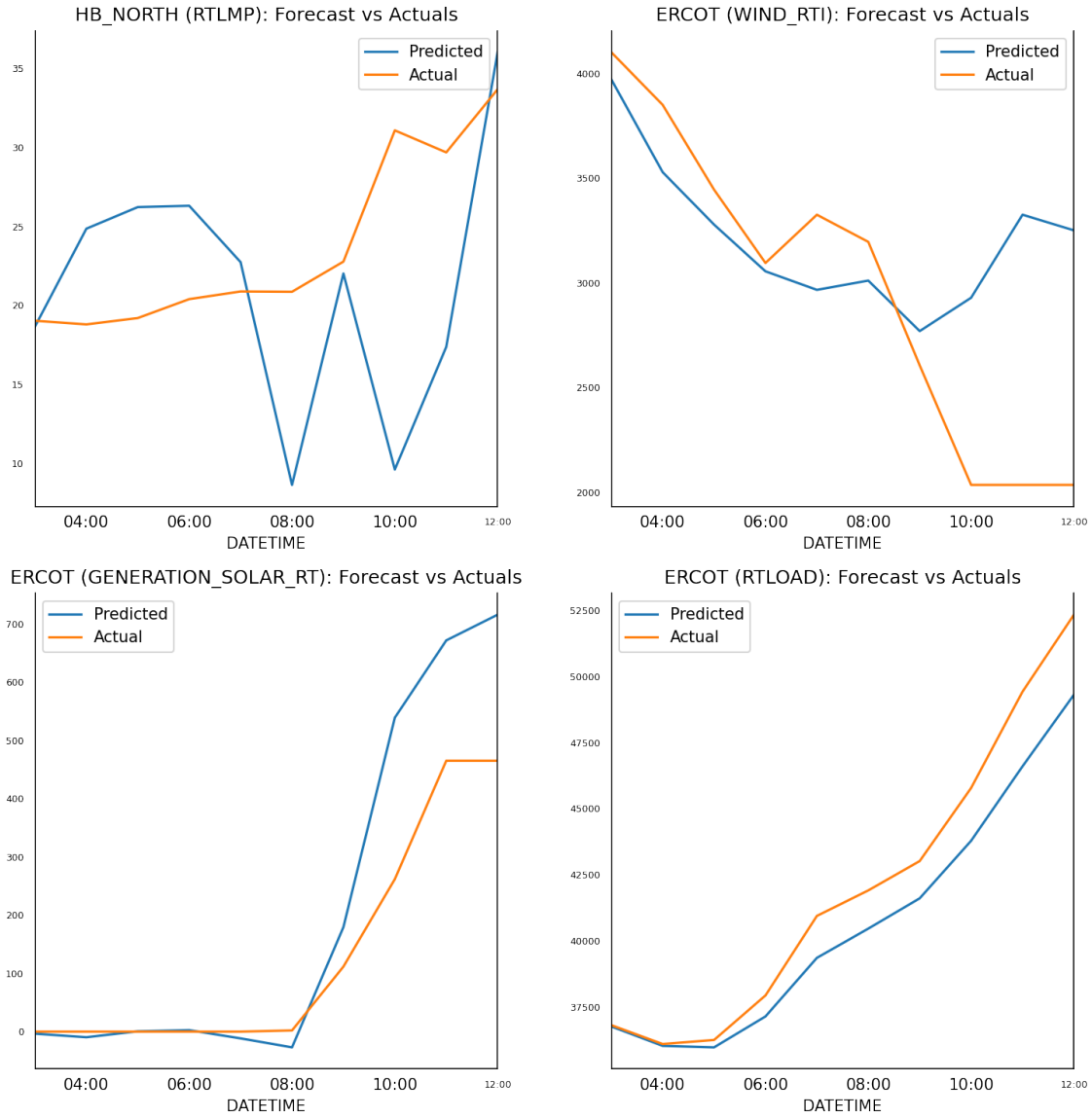


After fitting the model with the selected order number, checking for Serial Correlation of Residuals (Errors) using Durbin Watson test is vital, because we need to ensure there is no correlation left in the residuals. All values of this statistic are closed to 2, which means there is no significant serial correlation. My VAR model can perform 10 step prediction.

:

DATETIME	Predicted LMP	Actual LMP
2018-09-17 03:00:00	18.659732	19.0275
2018-09-17 04:00:00	24.847754	18.7975
2018-09-17 05:00:00	26.211239	19.1975
2018-09-17 06:00:00	26.295907	20.3950
2018-09-17 07:00:00	22.733143	20.8800
2018-09-17 08:00:00	8.654116	20.8600
2018-09-17 09:00:00	22.013849	22.7675
2018-09-17 10:00:00	9.626474	31.0600
2018-09-17 11:00:00	17.378585	29.6700
2018-09-17 12:00:00	36.003480	33.6607

My VAR model has good prediction for one step prediction, given the predicted LMP 18.659732 is closed to the actual LMP 19.0275 at 2018-09-17 03:00:00.
RMSE of LSTM is 9.41.



For wind generation, solar generation and RT load, the VAR is accurate to capture the trend.

3. **forecast model: LSTM**
   The first step is to split the data into train, val and test data sets. Then, I define the WindowGenerator to make sure I feed the training and validation data with the correct shape into the LSTM layer. I define adam optimization and MSE as the loss function. I create LSTM model, followed by a linear transformation layer. I train the LSTM model with 15 Epoches.

DATETIME	HB_NORTH (RTLMP)	Predict
2018-09-17 03:00:00	19.0275	19.230980
2018-09-17 04:00:00	18.7975	19.249119
2018-09-17 05:00:00	19.1975	19.220825
2018-09-17 06:00:00	20.3950	19.262356
2018-09-17 07:00:00	20.8800	19.240089
2018-09-17 08:00:00	20.8600	19.221453
2018-09-17 09:00:00	22.7675	19.154320
2018-09-17 10:00:00	31.0600	19.310219
2018-09-17 11:00:00	29.6700	19.265726
2018-09-17 12:00:00	33.6607	19.262032

My VAR model has good prediction for one-step prediction, given the predicted LMP 19.230980 is closed to the actual LMP 19.0275 at 2018-09-17 03:00:00.
RMSE of LSTM is 6.88.