Summary Report for Assignment 3

Henghui Qi

1. EDA

1.1 Dataset

The dataset has 14987 rows and 10 columns. Besides "DATETIME", which is of type "datetime", we care about "HB_NORTH (RTLMP)", "ERCOT (WIND_RTI)", "ERCOT (GENERATION_SOLAR_RT)", and "ERCOT (RTLOAD)", which are all float.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14987 entries, 0 to 14986
Data columns (total 10 columns):
DATETIME                        14987 non-null datetime64[ns]
HB_NORTH (RTLMP)                14987 non-null float64
ERCOT (WIND_RTI)                14982 non-null float64
ERCOT (GENERATION_SOLAR_RT)     14983 non-null float64
ERCOT (RTLOAD)                  14987 non-null float64
HOURENDING                      14987 non-null int64
MARKETDAY                       14987 non-null datetime64[ns]
PEAKTYPE                        14987 non-null object
MONTH                           14987 non-null object
YEAR                            14987 non-null int64
dtypes: datetime64[ns](2), float64(4), int64(2), object(2)
```

From the count of "PEAKTYPE", "WDPEAK" accounts for the biggest part, then "OFFPEAK", and "WEPEAK" has the smallest number. From the count of "MONTH", we can tell that the data is mainly from January to August.

```
WDPEAK      6966
OFFPEAK     4997
WEPEAK      3024
Name: PEAKTYPE, dtype: int64
```

```
AUGUST       1488
JANUARY      1488
MAY          1488
JULY         1488
MARCH        1486
JUNE         1440
APRIL        1440
FEBRUARY     1344
SEPTEMBER    1116
DECEMBER      744
OCTOBER       744
NOVEMBER      721
Name: MONTH, dtype: int64
```

From the summary of numeric data, "HB_NORTH (RTLMP)" and "ERCOT (WIND_RTI)" have relatively large standard deviations and imbalanced distributions.

	HB_NORTH (RTLMP)	ERCOT (WIND_RTI)	ERCOT (GENERATION_SOLAR_RT)	ERCOT (RTLOAD)	HOURENDING	YEAR
count	14987.000000	14982.000000	14983.000000	14987.000000	14987.000000	14987.000000
mean	25.766417	7532.436283	291.989714	42371.673703	12.495763	2017.415493
std	46.361945	3992.884834	370.914596	9874.339631	6.922309	0.492823
min	-17.860000	54.440000	0.000000	25566.511248	1.000000	2017.000000
25%	18.041250	4135.630000	0.000000	35431.636526	6.000000	2017.000000
50%	20.057500	7281.445000	22.150000	39934.007113	12.000000	2017.000000
75%	25.030000	10851.647500	608.635000	47873.100786	18.000000	2018.000000
max	2809.357500	20350.400000	1257.540000	73264.662123	24.000000	2018.000000

1.2 Duplicates and NaNs

There is no duplicate in the data.

```
df.iloc[:,1:5].duplicated().sum()#check duplicates
```

```
0
```

There are five NaNs in "ERCOT (WIND_RTI)" and four NaNs in "ERCOT (GENERATION_SOLAR_RT)".

```
HB_NORTH (RTLMP)              0
ERCOT (WIND_RTI)              5
ERCOT (GENERATION_SOLAR_RT)   4
ERCOT (RTLOAD)                0
dtype: int64
```

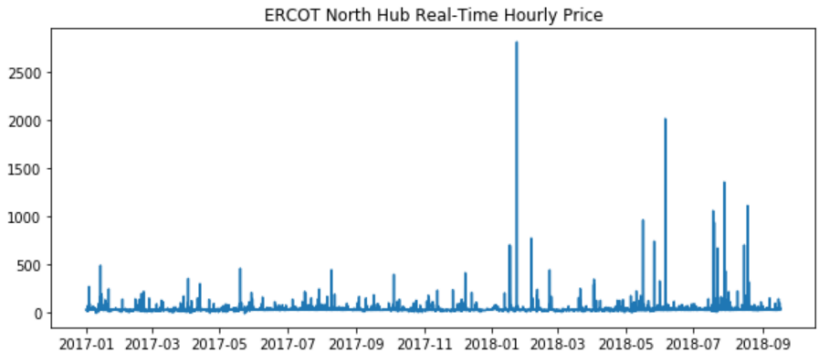
Fill NaN values with previous data and now there is no NaN left in the data.

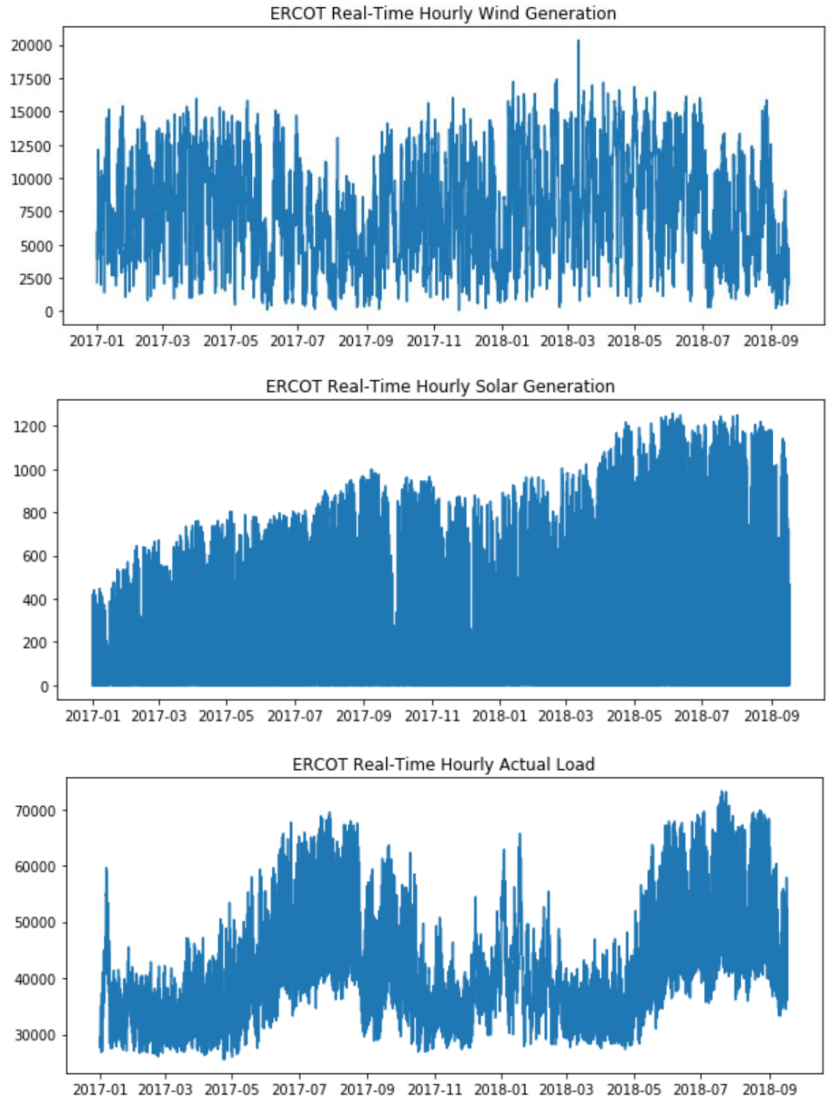
```
dff['ERCOT (WIND_RTI)'].interpolate(method='pad', inplace=True)#fill the missing value with the previous value
dff['ERCOT (GENERATION_SOLAR_RT)'].interpolate(method='pad', inplace=True)
'''for j in range(1,5):
    for i in range(dff.shape[0]):
        if(np.isnan(dff.iloc[i,j])):
            dff.iloc[i,j] = dff.iloc[i-1,j]'''
print(dff.iloc[:,1:5].isna().sum())
print(dff.iloc[:,1:5].isna().sum().sum())
```

```
HB_NORTH (RTLMP)              0
ERCOT (WIND_RTI)              0
ERCOT (GENERATION_SOLAR_RT)   0
ERCOT (RTLOAD)                0
dtype: int64
0
```

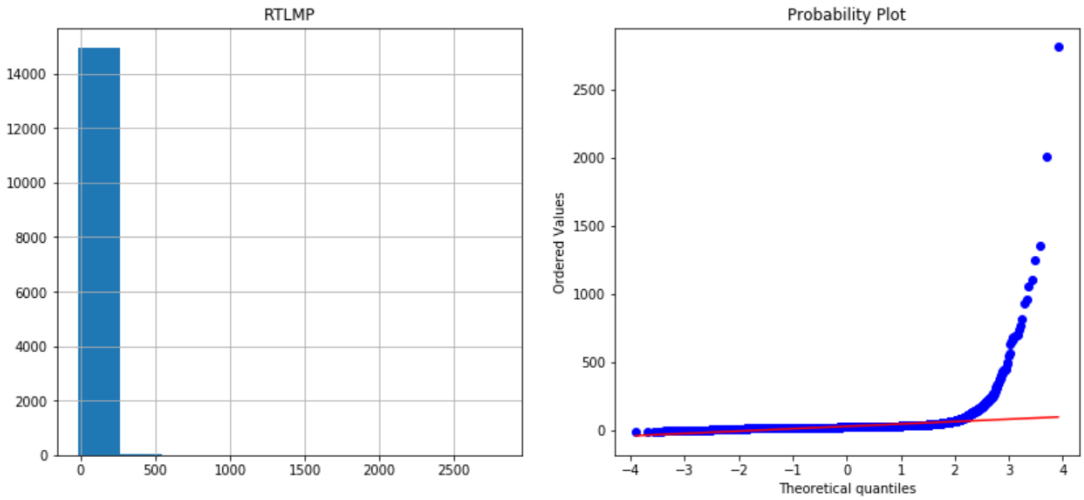
1.3 Plots

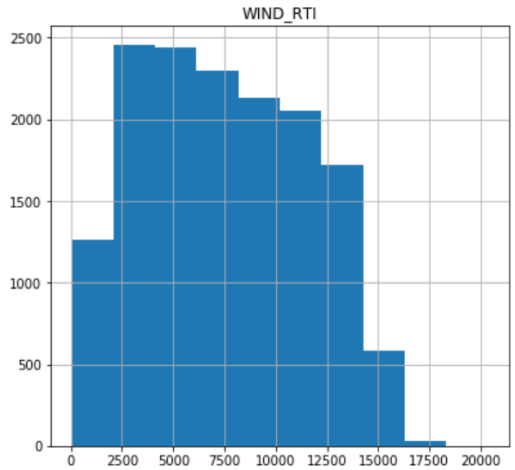
Plot the four variables we care about, with "DATETIME" as x-axis. Seasonal patterns are seen in these four plots.

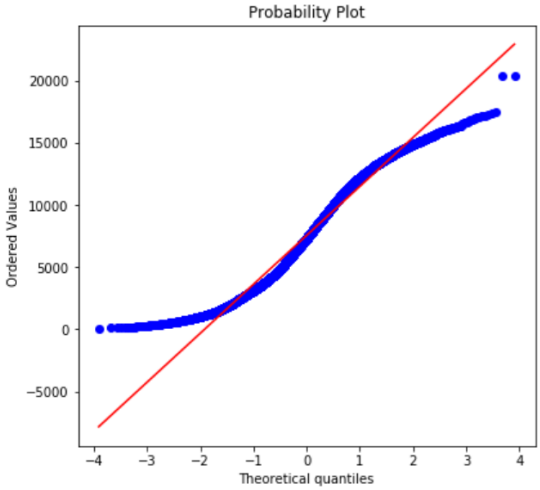


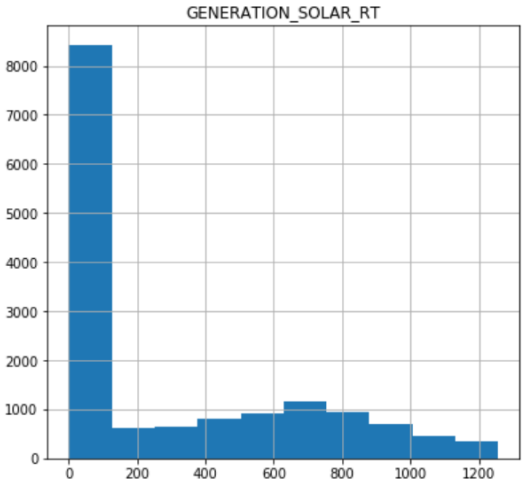


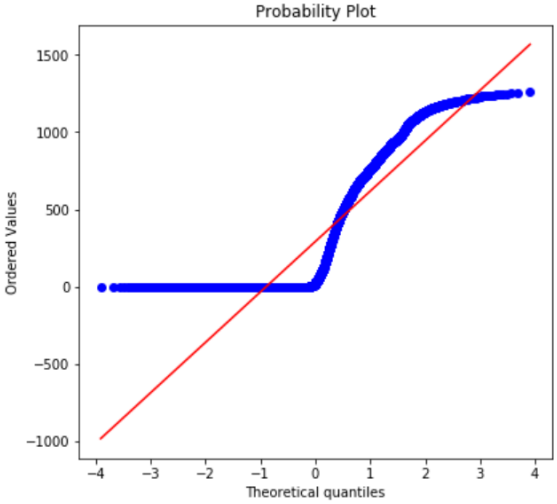
Histograms and normal probability plots of these four variables indicate the data is far from normally distributed, especially "HB_NORTH (RTLMP)" and "ERCOT (GENERATION_SOLAR_RT)".

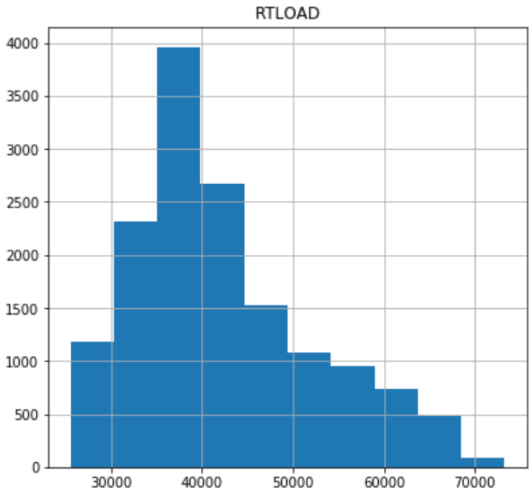


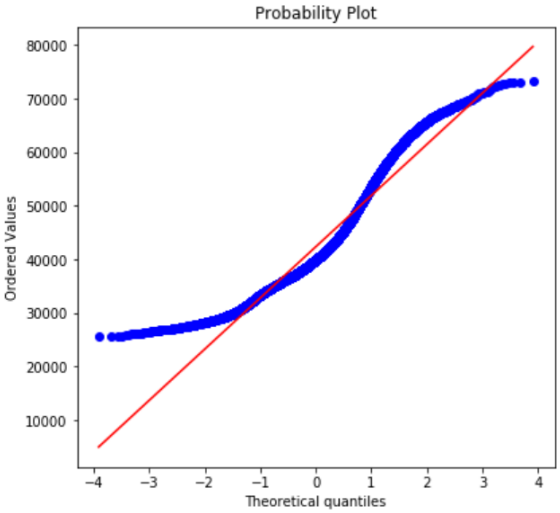

WIND_RTI
2500
2000
1500
1000
500
0
0
2500
5000
7500
10000
12500
15000
17500
20000


Probability Plot
Ordered Values
20000
15000
10000
5000
0
−5000
−4
−3
−2
−1
0
1
2
3
4
Theoretical quantiles

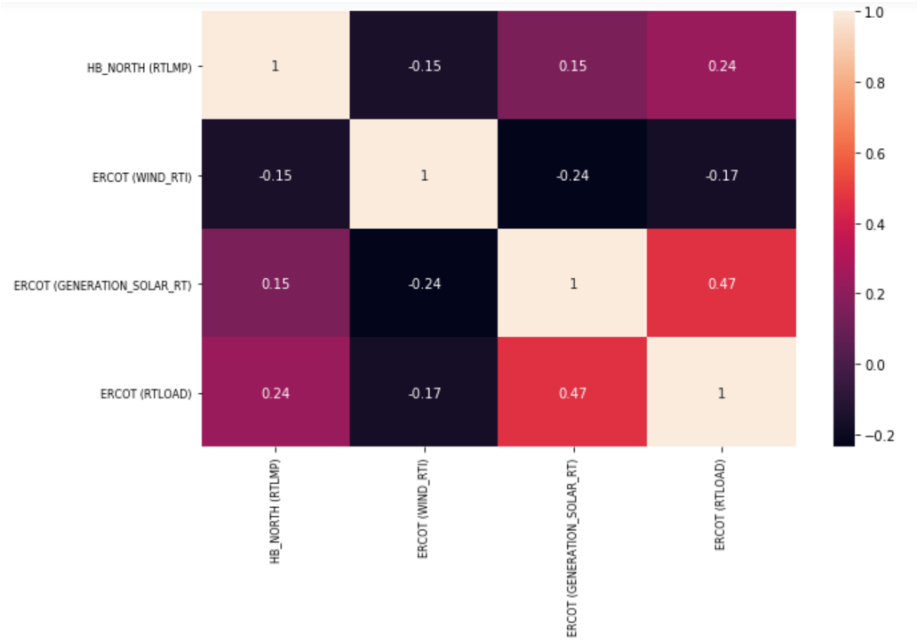

GENERATION_SOLAR_RT
8000
7000
6000
5000
4000
3000
2000
1000
0
0
200
400
600
800
1000
1200


Probability Plot
Ordered Values
1500
1000
500
0
−500
−1000
−4
−3
−2
−1
0
1
2
3
4
Theoretical quantiles


RTLOAD
4000
3500
3000
2500
2000
1500
1000
500
0
30000
40000
50000
60000
70000


Probability Plot
Ordered Values
80000
70000
60000
50000
40000
30000
20000
10000
−4
−3
−2
−1
0
1
2
3
4
Theoretical quantiles

The correlation plot of these four variables shows that they are not highly correlated. The largest correlation coefficient is 0.47 between “ERCOT (GENERATION_SOLAR_RT)”, and “ERCOT (RTLOAD)”.



2. Modeling and Prediction

2.1 Splitting Data

In order to test the prediction, dataset was split into “train” and “test”.

```
# create train and test data
train, test = data.iloc[:-24,:], data.iloc[-24:,:]#predict 24 hours
print(train.shape, test.shape)
```

```
(14963, 4) (24, 4)
```

2.2 Checking Before Fitting

2.2.1 Checking Stationarity

For Time Series modeling, data needs to be stationary. Use Augmented Dickey-Fuller (ADF) test to check whether data is stationary. From the result, all four variables are stationary series.

```
RTLMP:
Test Statistic          -1.828574e+01
p-value                  2.303118e-30
# Lags                   2.400000e+01
# Observations           1.493800e+04
Critical Value (1%)     -3.430788e+00
Critical Value (5%)     -2.861734e+00
Critical Value (10%)    -2.566873e+00
dtype: float64
 Series is Stationary
```

```
WIND_RTI:
Test Statistic          -1.158489e+01
p-value                  2.883379e-21
# Lags                   4.100000e+01
# Observations           1.492100e+04
Critical Value (1%)     -3.430788e+00
Critical Value (5%)     -2.861734e+00
Critical Value (10%)    -2.566873e+00
dtype: float64
 Series is Stationary
```

```
GENERATION_SOLAR_RT:
Test Statistic          -7.584895e+00
p-value                  2.622485e-11
# Lags                   4.200000e+01
# Observations           1.492000e+04
Critical Value (1%)     -3.430788e+00
Critical Value (5%)     -2.861734e+00
Critical Value (10%)    -2.566873e+00
dtype: float64
 Series is Stationary
```

```
RTLOAD:
Test Statistic             -5.077353
p-value                     0.000016
# Lags                     42.000000
# Observations          14920.000000
Critical Value (1%)        -3.430788
Critical Value (5%)        -2.861734
Critical Value (10%)       -2.566873
dtype: float64
 Series is Stationary
```

2.2.2 Checking Causality

Use Granger Causality Test to investigate causality of data. From the result that all values are all below 0.05 except the diagonal, we can reject the null hypothesis that x does not cause y.

	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

2.3 Building the model

Vector Auto Regression (VAR) is one of the commonly used methods for multivariate time series forecasting. In a VAR model, each variable is a linear function of the past values of itself and the past values of all the other variables, which enables it to understand and use the relationship between several variables. This is useful for describing the dynamic behavior of the data and it also provides better forecasting results. Here I fitted the VAR model on the train and the summary is as shown below. The AIC is 41.3467. The biggest correlations are 0.072 ("HB_NORTH (RTLMP)" & "ERCOT (RTLOAD)") and 0.055 ("ERCOT (WIND_RTI)" & "ERCOT (RTLOAD)", which are small enough to ignore in this case.

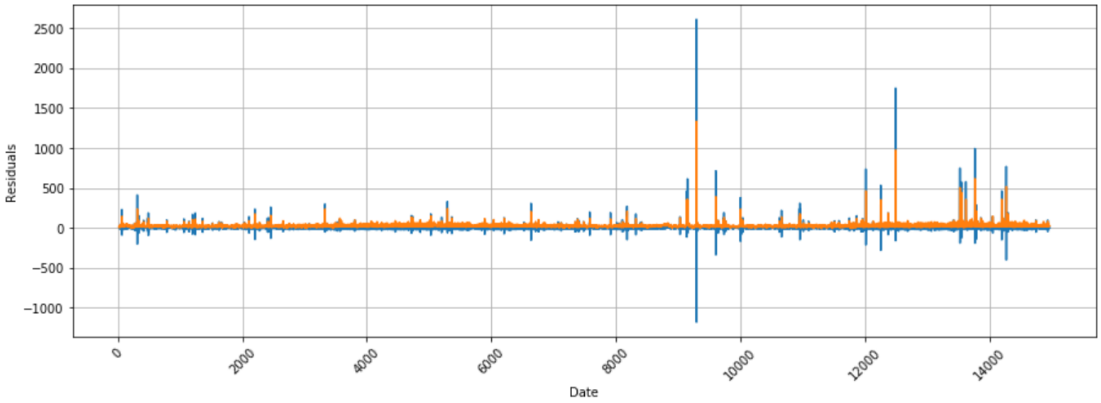
```
  Summary of Regression Results
==================================
Model:                         VAR
Method:                        OLS
Date:           Mon, 16, May, 2022
Time:                     12:52:57
----------------------------------------------------------------------
No. of Equations:         4.00000    BIC:                    41.5117
Nobs:                     14943.0    HQIC:                   41.4014
Log likelihood:          -393410.    FPE:                 9.04956e+17
AIC:                      41.3467    Det(Omega_mle):      8.85597e+17
----------------------------------------------------------------------
```

Correlation matrix of residuals

	HB_NORTH (RTLMP)	ERCOT (WIND_RTI)	ERCOT (GENERATION_SOLAR_RT)	ERCOT (RTLOAD)
HB_NORTH (RTLMP)	1.000000	-0.052043	-0.003273	0.072442
ERCOT (WIND_RTI)	-0.052043	1.000000	-0.003859	0.054809
ERCOT (GENERATION_SOLAR_RT)	-0.003273	-0.003859	1.000000	0.117179
ERCOT (RTLOAD)	0.072442	0.054809	0.117179	1.000000

2.4 Plotting and Testing the Residuals

Plot the residuals of "HB_NORTH (RTLMP)" against fitted values of "HB_NORTH (RTLMP)". The residual plot looks normal with constant mean.



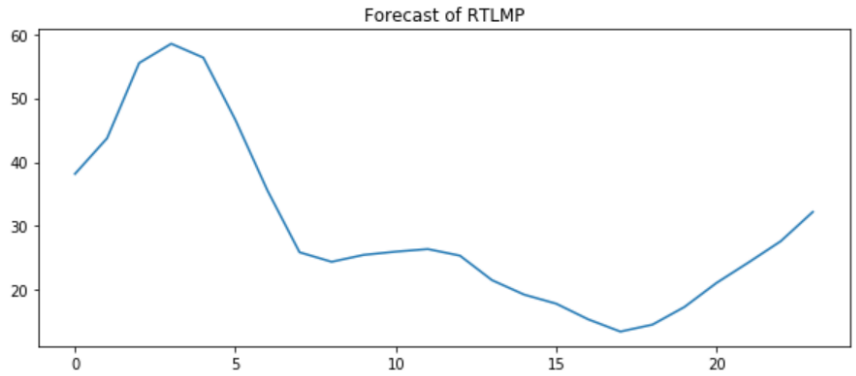
Do the Durbin-Watson statistic test to see whether there is a correlation in the residuals in the fitted result. The result of residuals in the model are around 2, between 1.5 and 2.5, which means autocorrelation is likely not a cause for concern.

```
HB_NORTH (RTLMP) : 2.0011419602169327
ERCOT (WIND_RTI) : 2.000349301082866
ERCOT (GENERATION_SOLAR_RT) : 2.1028136039062484
ERCOT (RTLOAD) : 1.9975613354295998
```

2.5 Predicting and Evaluating

Predict the next 24 hours data. The data frame of first five hours prediction of four variables and the plot of 24 hours prediction of "HB_NORTH (RTLMP)" are shown as below.

	HB_NORTH (RTLMP)	ERCOT (WIND_RTI)	ERCOT (GENERATION_SOLAR_RT)	ERCOT (RTLOAD)
0	38.154912	2041.531283	493.162012	51483.342245
1	43.729823	2107.651186	474.903818	53837.209191
2	55.497889	2316.037021	503.228542	55347.101220
3	58.509814	2725.733864	551.013823	55944.077449
4	56.336830	3156.737245	570.187119	55502.552441



The bias of prediction of "HB_NORTH (RTLMP)" is 2.473 and the RMSE of it is 8.808. The bias and RMSE are low here, indicating the right fit of the model is obtained.

```
Bias: 2.472717
RMSE: 8.808111851099788
```