EDA:

- Check missing values

We only have 9 missing values, it is safe to drop rows containing missing values.

- Check duplicates

No duplicates.

- Check correlations between variables

The dependent variable RTLMP does not show strong correlation with any initial independent variable.

The good news is no pair of explanatory variables shows high correlation. We will avoid multicollinearity problem.

- Check time series continuity

Three discontinuous points are detected with a gap of 2 to 3 hours. Fill the gap with previous valid observation.

- Scatter plot between independent variable and independent variable

The scatter plots ensure the idea that they have weak correlation.

- Outliers

Draw the box plot of RTLMP and find that around 50% of observations fall out of the box.

It seems that using the quantile method to determine outliers is not suitable for RTLMP. Further study could be on how to decide the outliers for RTLMP.

The good news is I am going to use Random forest for modeling and this model is not sensitive to outliers.

Feature Engineering:

- Convert 'PEAKTYPE' to binary type feature
- Convert 'MONTH' to numerical data
- Make hour of day (HOURENDING) cyclic
- Add lag terms of RTLMP, WIND_RTI, GENERATION_SOLAR_RT, RTLOAD
  e.g. RTLMP_1hour_lag, RTLMP_6hour_lag, etc.

After feature engineering, we have 30 explanatory variables in total.

Modeling:

We select RMSE as metric.

We use 75% for training and the rest for testing.

- Baseline model

The baseline model is to use previous hour's RTLMP as prediction.

RMSE on test set: 38.28

- Random Forest

1. Train the model with all the features we have (30 in total).

2. Get the feature importance for each feature and select the top 10 most important features.

3. Retrain the model with the selected features.
   Parameters like max depth, max features and number of features to be selected from the first model are selected based on cross validation.

   RMSE on test set:34.58

Feature importance:

	importance
HB_NORTH (RTLMP)_1hours_lag	0.291595
ERCOT (GENERATION_SOLAR_RT)	0.213062
ERCOT (RTLOAD)_168hours_lag	0.073174
ERCOT (GENERATION_SOLAR_RT)_72hours_lag	0.071507
ERCOT (RTLOAD)	0.065356
HB_NORTH (RTLMP)_24hours_lag	0.058312
HB_NORTH (RTLMP)_12hours_lag	0.058282
ERCOT (WIND_RTI)_1hours_lag	0.058005
ERCOT (RTLOAD)_1hours_lag	0.055830
ERCOT (RTLOAD)_6hours_lag	0.054875

It is interesting to see the hour of day and the peak type don't have much power for explaining RTLMP. And for WIND_RTI, only the one-hour lag value is important.