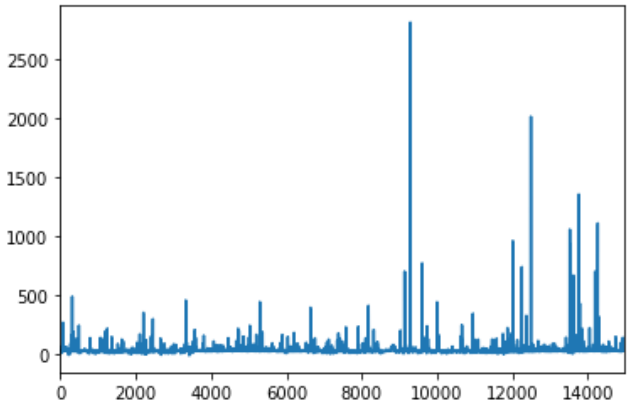
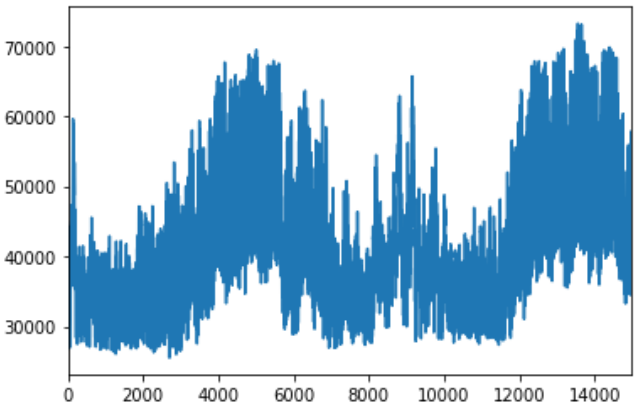
Assignment 3 Report

In this assignment, I conducted exploratory data analysis, feature engineering and model development.

In EDA, the data set is firstly loaded and applied the dropna() function to drop the rows with nan values. I then inspected the data set and visualized the target and three numeric and non-date time variables. I observed that there exists cyclicity and seasonality in the time series plots. Example plots:



Target variable (ERCOT North hub real-time price)



Independent variable (actual load)

In the second part, I engineered PEAKTYPE, month and hour into new features. PEAKTYPE has three values and it can be transformed into dummy variables using one-hot encoding. For month, considering that the price has seasonality in it, I first transform the 12 months into 4 quarters and then used one-hot encoding. For day and hour, I transformed each variable with a pair of variables with the sin() and cos() function, so that to make sure that days like begin and end of a month have similar values, and hours like 0 and 24 have similar values.
Data set after feature engineering:

	ERCOT (WIND_RTI)	ERCOT (GENERATION_SOLAR_RT)	ERCOT (RTLOAD)	peaktype_OFFPEAK	peaktype_WDPEAK	Q_1	Q_2	Q_3	sin_day	cos_day	sin_hour	cos_hour
0	2155.31	0.0	29485.791355	1	0	1	0	0	0.207912	0.978148	0.258819	0.965926
1	2313.81	0.0	28911.565913	1	0	1	0	0	0.207912	0.978148	0.500000	0.866025
2	2587.68	0.0	28238.258175	1	0	1	0	0	0.207912	0.978148	0.707107	0.707107
3	2748.65	0.0	27821.000513	1	0	1	0	0	0.207912	0.978148	0.866025	0.500000
4	2757.49	0.0	27646.942413	1	0	1	0	0	0.207912	0.978148	0.965926	0.258819

For the last part, model development, I split the data set into training and test sets and tried 2 regression models, the ridge regression and the random forest regression. Then I compared the two models with RMSE and found the random forest model has better performance. The RMSE for the two models are 38.89 and 30.99 respectively. This section can be improved by adding a hyperparameter tunning part with Gridsearch or more complexed models.