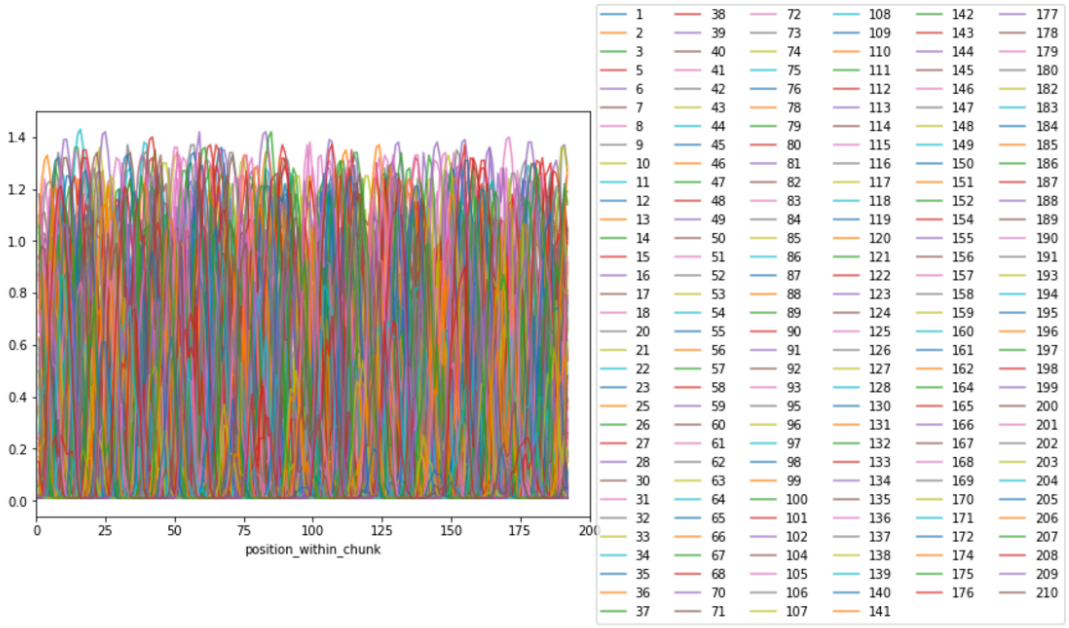
Capstone Project 1 – Data Wrangling

The training data set for the project includes hourly data points for 8 days from various sites around Chicago. Each "chunk" of data includes 192 data points that start at various time of day, day of week, and months of the year. So by plotting the solar radiation sequentially by position within each chunk, we get the following below.

Solar Radiation



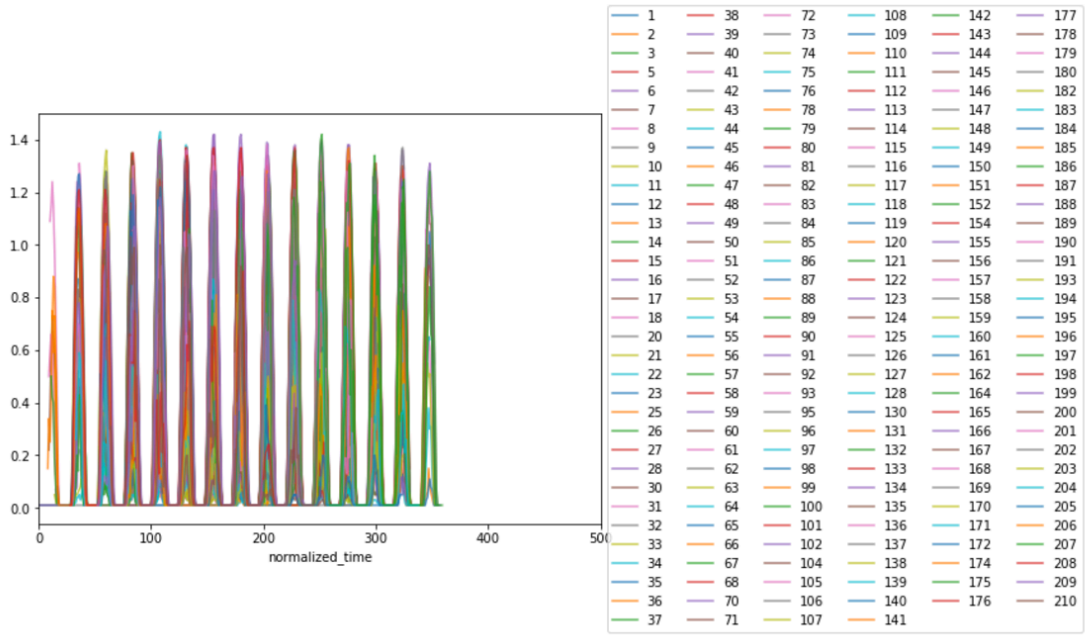
Because the time series were not synced for time of day and day of the week, it would not be possible to see the effect of due to time of day or day of week effects, if any. So in order to standardize, I put each chuck of data on a standardized time line that starts at midnight of Sunday. Each chunk of data will fall in the somewhere in the new standardized time.

Standardized Time Conversion:

Standardized	0	24	48	72	96	120	144	168	192	216	240	264	288	312	336	360	384
Hour	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00
Day	Sun	Mon	Tue	Wed	Thr	Fri	Sat	Sun	Sun	Mon	Tue	Wed	Thr	Fri	Sat	Sun	Mon

Replotting the solar radiation by the new standardized time, one can see the data makes a lot more sense as the daily cycle of solar radiation correlates very well to time of day.

Radiation Data using new standardized time.



Note: There are 2 chunks missing in the training data (#94 and #192). According to Kaggle, there are many rows for which some of the measurements are missing for both the training and evaluation data. It was intended to ignore those in calculating MAE score. We were told to transformed all NA's to "-1000000".