DATA CLEANING

- I replaced the missing values in “ _conds” column by the most frequently occuring value in the week before and after. The logic behind it is that weather condition tend to be similar over a short span of time.

- The missing values of " _dewptm", “ _hum”, " _pressurem", “ _tempm”, “ _vsim”, " _wdird" and " _wspdm" are filled by the average of the previous day and the next day value as these weather conditions tend to increase or decrease gradually for most part.

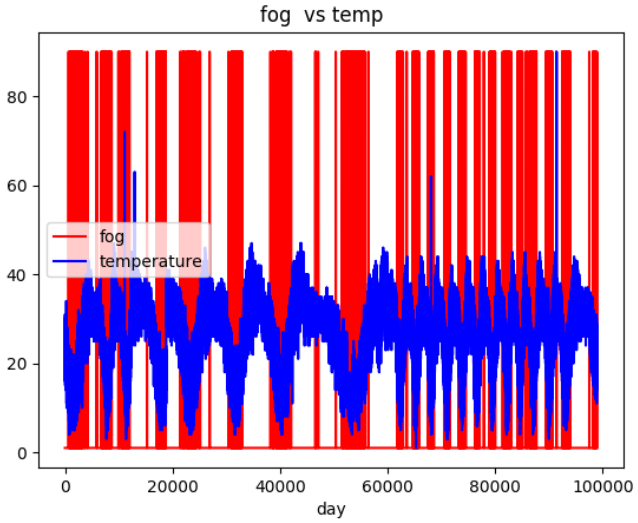
- The missing value of “ _wdire” is replaced with it’s previous day’s value as wind direction tend to be similar on a seasonal basis.

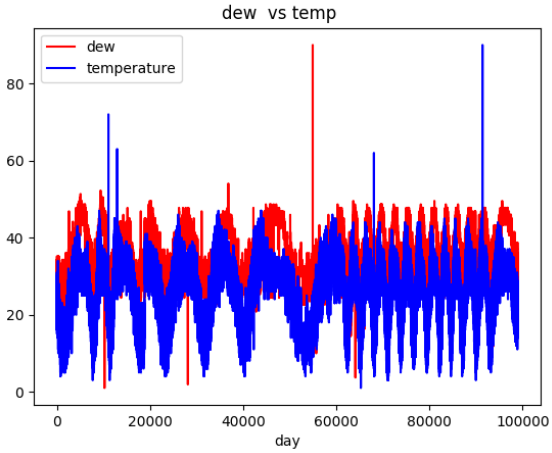
- The categorical values in the columns " _conds", " _wdire" is changed to an int type by doing one hot encoding.

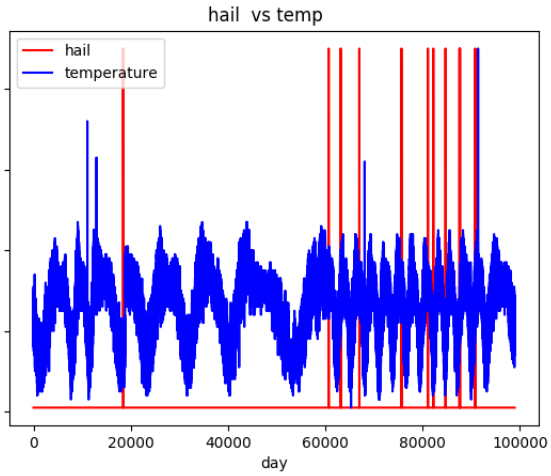
- Next, I used MinMaxScaler to scale all the columns to a same range. I however did not scale it to (0,1) and opted to go for the range of the “ _tempm” column, so that I don’t have to reverse scale the predicted output.

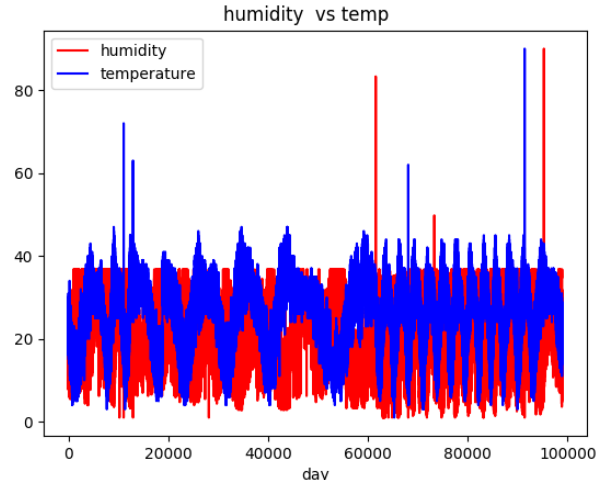
DATA ANALYSIS

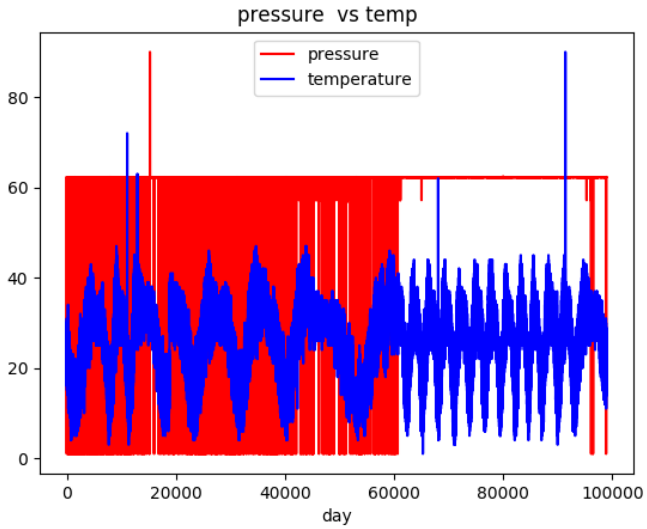
- I plotted the features against the temperature. And then used the same to derive my inference.

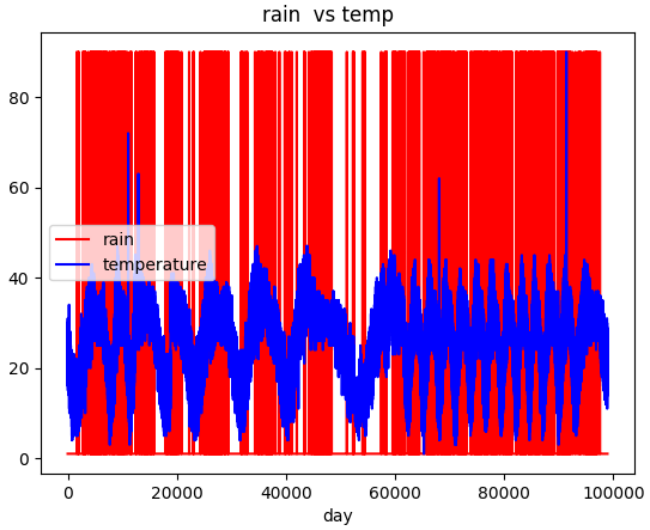


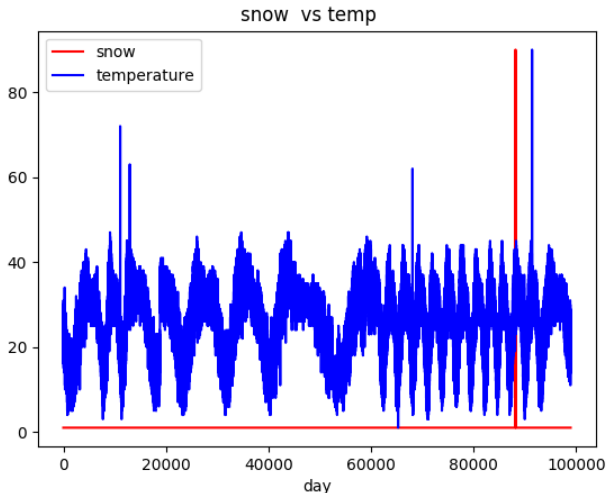


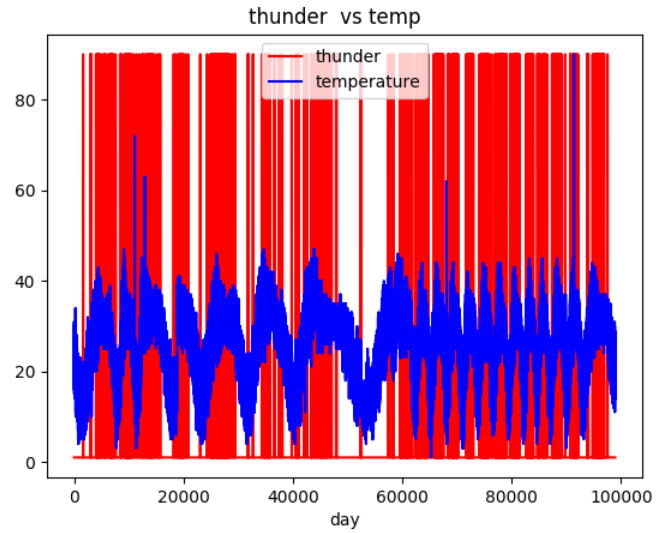

hail vs temp
hail
temperature
day

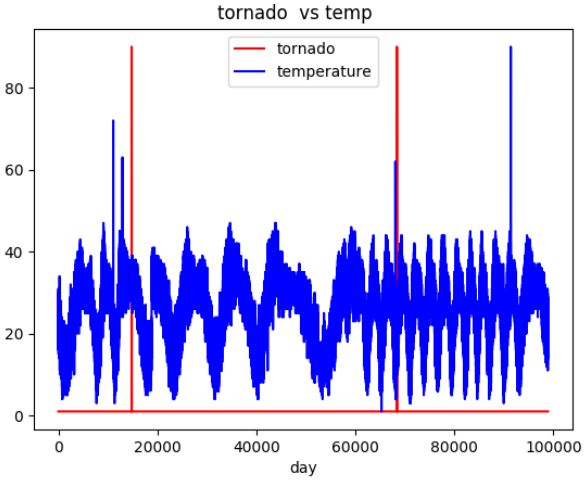

humidity vs temp
humidity
temperature
day

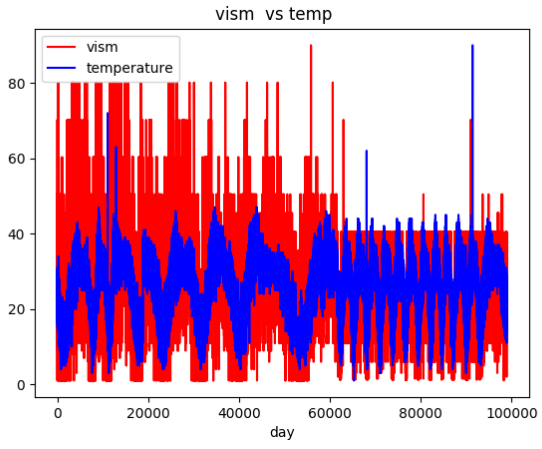

pressure vs temp
pressure
temperature
day

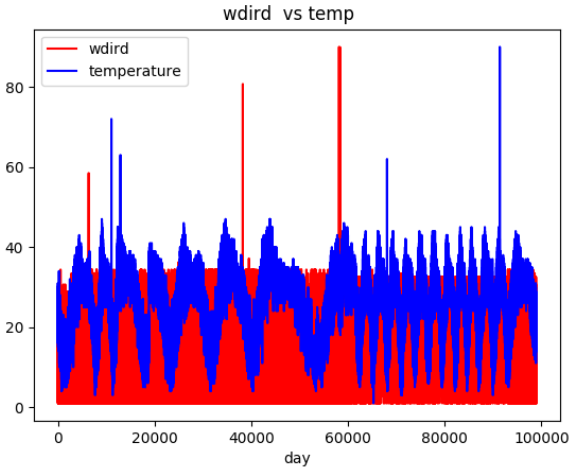

rain vs temp
rain
temperature
day

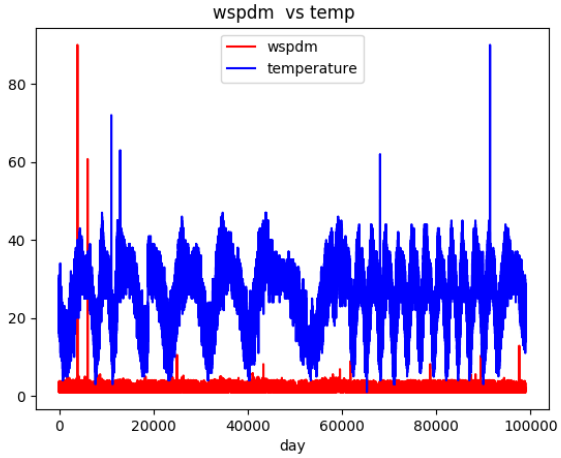

snow vs temp
snow
temperature
day

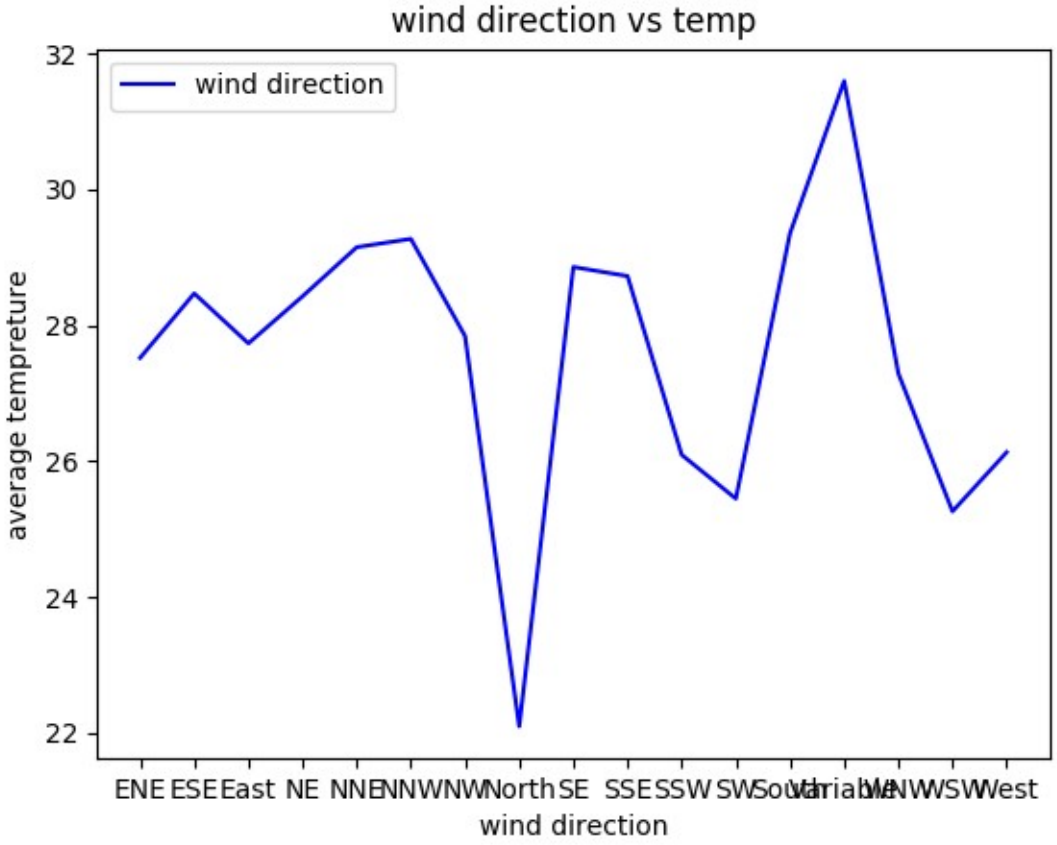

thunder vs temp
thunder
temperature
day


tornado vs temp
tornado
temperature
day


vism vs temp
vism
temperature
day







- Fog- inversely proportional temperature.
- Dew – directly proportional to temperature
- Hail – Doesn’t seem to have a direct correlation to temperature.
- Humidity - inversely proportional temperature.
- Pressure – A consistent high pressure seems to increase the frequency of change of temperature.
- Rain - directly proportional temperature.
- Snow - Doesn’t seem to have a direct correlation to temperature.
- Thunder - directly proportional temperature.
- Tornado - Doesn’t seem to have a direct correlation to temperature.

- Vsim – Is high for a slower changing temperature.
- Wdird - Doesn’t seem to have a direct correlation to temperature.
- Wspdm - Doesn’t seem to have a direct correlation to temperature.
- Wdire – Clearly dependent.

DATASET AND ALGORITHM

For the purpose of time series forcasting, I have taken into consideration previous 360 days as it approximately encompasses the previous year’s weather condition. I have trained an LSTM model with 2 datasets.

1. With just the temperature of the previous 360 days.

2. Along with the temperature of the previous 360 day, I have considered the other features other than : "datetime_utc", " _hail", " _snow", " _tornado", " _wdird", " _wspdm".

The LSTM model is very simple and has 2 hidden layers of 50 nodes each. Followed by a fully connected layer with one output unit.

1. The multivariate dataset result:

2. The dataset with only temperature: