

Question 1: How to deal with the null values in the data pre-processing process?

First, we check if the entire column contains only N/A values. In this case, we delete this column. Then, we delete the rows that still have N/A values.

Question 2: What year does the data come from?

The training data are from 2016. The test data are from 2017 and 2018. We use one-year data to predict two-year data.

Question 3: Is there anything special to do to deal with such a large amount of data?

We use hourly data to calculate daily average data to reduce the data size.

Question 4: Will changing from hourly to daily readings lose any information?

This approach may cause some information loss. For example, when we convert the weather data from hourly data to daily data, we cannot get the temperature difference between noon and night. However, because our prediction focuses on the data of a whole year, we can allow some loss of information like this. Since we want to predict the data of a whole year, we care more about the trend of a whole year than the trend of every hour.