Readme

There are mainly three notebooks in the ‘src’ directory. They are listed and described below:

1. Data_Preprocessing.ipynb
   This notebook is maily for data preprocessing. It will generate the preprocessed data files in two directories named ‘cleaned_data’ and ‘imputed_data’.
   - The ‘cleaned_data’ will include the data files with NaN value for the 35 air quality stations.
   - The ‘imputed_data’ directory includes the One-Hot encoded and data imputed files for final prediction.

2. XGBoost.ipynb
   This notebook will generate the final ‘submission.csv’ file with the model XGBoost.

3. LSTM.ipynb
   This notebook will use the LSTM model for data analysis and simple prediction. Since the result of LSTM is not better then XGBoost. So this notebook will not make the final prediction for the 35 air quality stations.

Meanwhile, there is a directory name ‘MSBD5002PROJECT_data’ in the ‘src’ directory. There are three files in this directory:

- **airQuality_station.csv**: includes the station location information for 35 air quality stations
- **Beijing_grid_weather_station.csv**: includes the location information for the grid weather stations
- **observed_weather_stations.csv**: include the location information for the observed_weather stations