PROGRAMMING QUESTION

PLEASE READ BELOW CAREFULLY

- Please read and implement the below problem in Jupyter Notebook environment. Your final notebook code should be presenting your ideas and results in a concise way.

- After finishing the task, you should upload the notebook (and other essential things if there are) to your favourite git provider (github, bitbucket, gitlab ... etc.) as a publicly accessible code repository.

- Paste the URL of the solution repository to your solutions document as answer to 17th question.

- If you are selected to proceed to technical interview, you will be responsible to explain and discuss your algorithm in detail on notebook.

- If you think there is an ambiguity in the task from your point of view, please write down the ambiguity, the cause of it and your reasoning to accept your assumption to solve this ambiguity in the notebook.

TASK: Forecasting bus demand in Banana Republic municipalities.

The central urban planning commitee of Banana Republic asked you to help them with the forecast of bus demands of municipalities. And they provide a nice dataset to support you (https://pi.works/3w8IJbV).

The dataset includes two measurements for an hour for the number of used buses in each municipality, each measurement is timestamped. The dataset format is as follows (comma separated values):

MUNICIPALITY\_ID, TIMESTAMP, USAGE, TOTAL\_CAPACITY

where municipality\_id is an anonymization to disguise the actual names, timestamp represents the exact time of the measurement, usage is the number of buses in use at the time of measurement and total\_capacity represents the total number of buses in the municipality. There are 10 municipalities (ids from 0 to 9), and two measurements for an hour.

The committee asks you to forecast the hourly bus usages for next week for each municipality. Hence you can aggregate the two measurements for an hour by taking the max value (sum would not be a nice idea for the obvious reasons) for each hour, and you should model this data with a time series model of your selection. (It would be a nice idea to implement a very simple baseline model first, and then try to improve the accuracy by introducing more complex methods eventually. The bare minimum requirement of the task is one simple baseline and one complex method.)

The committee says that they will use the last two weeks (starting from 2017-08-05 to 2017-08-19) as assessment (test) data, hence your code should report the error (in the criterion you chose for the task) for the last two weeks. You may use true values for the prediction of the last week of test data, then combine the error of the first and last week of the test separately.

Keep in mind that the dataset has missing data, hence a suitable missing data interpolation would be useful.