Instructions: In this project, you are given a dataset collected by an actual IoT system (see description below) and asked to use the dataset to build a forecasting model. You have to answer a set of questions, as well as propose your own interesting questions. The project is worth 40 marks.

1. Form teams of 4 students each and register your team in LumiNUS under Class Groups.
2. Do the following for each question. Use a Jupyter notebook (ipynb file) to do the analysis and answer all the parts of the question. Use both code cells (for code) and markdown cells (for comments). Submit (i) PDF file/Print preview of your Jupyter notebook, (ii) the Jupyter notebook (ipynb file), and (iii) any additional data you used. Do not include the original data I gave you. Zip all files into one zip file, name it apprpriately, and upload it to the correct LumiNUS folder.
3. Complete all of Question 1. Please name your file GroupName_Q1.zip and upload to LumiNUS. (10 marks)
4. Complete all of Question 2. Please name your file GroupName_Q2.zip and upload to LumiNUS. (10 marks)
5. Complete Question 3. For Question 3, please also include a detailed description of your proposed work. Be sure to justify all assumptions, give the necessary details, and make a persuasive argument for your proposal. Please name your file GroupName_Q3.zip and upload to LumiNUS. (10 marks)
6. Your presentation will be in the form of a video. The video should be 8-10 minutes long and focus on Question 3. Be sure to tell an interesting story and make a convincing argument. You can present results from Q1 and Q2 if they support what you want to do in Q3. It would be good if all members could play a role in the video. (10 marks)

Data Description: The data is available in the project directory in LumiNUS Files. In this project, we will consider natural gas consumption data from residential consumers. The smart gas meter data used for this paper was obtained from the Pecan Street project (https://www.pecanstreet.org/). The source of the data are homes in the Mueller neighborhood of Austin, Texas, USA. The homes in this neighborhood are primarily newly constructed, and include single-family homes, apartments, and town homes. Itron Centron SR smart gas meters are deployed in these homes and these meters send their information to a gateway inside the home. The gateway uses the home's Internet connection to send the data to the meter data management system (MDMS) or the processing center. The gas meters measure the cumulative gas consumption at a frequency of 15 seconds. The meters report a reading (in terms of the cumulative consumption) when the last marginal 2 cubic foot (or higher) of natural gas passes through the meter. Data from a six month interval (1 Oct 2015 to 31 Mar 2016) has been provided. The data has the following format:

```
<Timestamp (localtime)> <MeterID (dataid)> <meter reading (meter_value)>
```

The timestamp provides the date as well as the the hour and minute values when each reading was taken. Each meter has an unique identifier (MeterID). Recall that the meter readings are cumulative and not generated at periodic intervals.

Questions:

1. Exploring the Data (10 marks)

   1.1 How many houses are included in the measurement study? Are there any malfunctioning meters? If so, identify them and the time periods where they were malfunctioning. The information below regarding data collection may be useful.

1.2 Generate hourly readings from the raw data. Select one month from the 6-month study interval and plot the hourly readings (time-series) for that month. Hint: You will have to decide what to do if there are no readings for a certain hour.

1.3 Intuitively, we expect that gas consumption from different homes to be correlated. For example, many homes would experience higher consumption levels in the evening when meals are cooked. For each home, find the top five homes with which it shows the highest correlation.

2. Forecasting (10 marks)

2.1 In this part, you will asked to build a model to forecast the hourly readings in the future (next hour). Can you explain why you may want to forecast the gas consumption in the future? Who would find this information valuable? What can you do if you have a good forecasting model?

2.2 Build a linear regression model to forecast the hourly readings in the future (next hour). Generate two plots: (i) Time series plot of the actual and predicted hourly meter readings and (ii) Scatter plot of actual vs predicted meter readings (along with the line showing how good the fit is).

2.3 Do the same as Question 2.2 above but use support vector regression (SVR).

3. Student Proposal (10 marks)

3.1 At this point, you understand the data quite well. Propose and carry out additional analysis using the dataset given. Please be sure to justify why this additional analysis is useful and interesting.

Additional Information about Data Collection:

1. Gas flow meters have a sensor that is used to measure the volume of gas that passes though a pipe. Different meters use different sensors (e.g. ultrasonic sensors, synthetic diaphragm with rotating valve etc.). The meters check on the sensors periodically to get a reading of the current consumption value. This is what is meant in the sentence above: "The gas meters measure the cumulative gas consumption at a frequency of 15 seconds."
2. Now, just because the meter has obtained a reading from the sensors, it does not not have to send the reading off to the meter data management system (MDMS). Imagine 1.3 million households in Singapore sending out gas readings every 15 seconds to Singapore Power. The processing and bandwidth requirements may be too high for Singapore Power. So Singapore Power may wish for the meters to report at a lower frequency or when the consumption exceeds a certain threshold. However, the smart meter manufacturer does not know what is the reporting criterion of its users. So it builds meters that can read every 15 seconds because it thinks that this is a frequency that is high enough for all potential customers. The "reporting" frequency to the MDMS (as opposed to the "measuring" frequency) can be determined by the user of the meter such as Singapore Power.
3. So when are the meters supposed to "report" to the MDMS? The documentation that came with the data says "once the marginal consumption exceed 2 cubic meters". As you may observe in the data, this is not necessarily the case in some of the readings. So is that an anomaly? That is for you to decide and justify. If you were Singapore Power, under what circumstances would you think that a meter reading is suspicious and decide to investigate? Remember that there are two sides to the story. If you do not receive a reading from a meter for a really long time, would you think that the meter is defective? So would that justify sending a reading even if the consumption has not increased?