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.

1. Form teams in groups of 4 students and select a name for your team. Be creative! Please email us your group members and team name by Monday 15 Oct 2018, 12 noon.
2. Start with the first set of questions (Interim). Please use a Jupyter notebook (ipynb file) to do the analysis and write the report. Generate a PDF file of your Python notebook. Upload both your ipynb and PDF files to the IVLE workbin inside the appropriate folder by the due date (Monday 29 Oct 2018, 11:59PM)
3. Then, extend your Interim file to work on the second set and third set of questions. Again, upload your Final ipynb and PDF files to the IVLE workbin inside the appropriate folder by the due date (Monday 12 Nov 2018, 11:59PM)
4. Please name your files as Group_Name_Interim.ipynb and Group_Name_Final.ipynb. The same goes for the PDF files.

Data File:

The data file is available in the IVLE workbin under the directory ”Project Details”.

Data Description:

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 (October 1, 2015 to March 31, 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:(40 pts)

Use the data to answer the following in a Jupyter notebook.

1. Interim (10 pts)

   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.

   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.

   1.4 In Question 2 below, you will have to analyze the data further. In addition to what you are asked to do below, please propose additional analysis you can do with the data. Justify why you would carry out the analysis. If you give us your suggestions at the Interim, we will give you feedback.

2. Forecasting (20 pts)

   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 pts)

   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.

Presentation:(10 pts)

We will have a presentation session in which each group will present their findings. Details on the presentation will be given later.