EM–CURE Assignment
***E**ntrepreneurially **M**inded **C**ourse-based **U**ndergraduate **R**esearch **E**xperiences*

Course: ECE 4610, Electrical Energy Conversion
Due: May 10, 2024

What is the purpose of this assignment?

The purpose of this EM-CURE assignment is to practice two critical components of the research process: data processing and data analysis. These components can occur at any stage in the research process. In electrical engineering, before conducting an experiment or running a simulation, you might have a dataset from a third party that requires verification and comprehensive understanding before its use. This verification and understanding involve ensuring the data is error-free and presenting it visually to enhance comprehension of its content. Similarly, after performing the experiment or running the simulation, the collected data must be verified for the veracity of the experimental testbed (e.g., properly calibrated sensors) or the functionality of the simulation model (e.g., no coding errors in your model).

This assignment will guide you through the processes of *data verification* (processing) and understanding (visualization) for the annual energy consumption of three buildings.

What data will you process and plot?

The data for this assignment comes from a National Renewable Energy Laboratory database containing load profiles (energy consumption curves) of different buildings. Because this dataset represents the different uses, kinds, and behaviors of buildings in the commercial sector, a lot of research has been conducted regarding the energy (e.g. sustainable energy integration, energy reliability analysis…).

I've placed the necessary data in this Google Drive location. The folder contains data for three buildings (large hotel, office, and full-service restaurant), with each building having 30 features (columns). The first 5 are 'identifiers' that describe the dataset, while the last 25 consist of metered data. The metered data encompasses various aspects, from natural gas used by cooking stoves to exterior building lighting. All 25 metered data columns are reported in kWh. Note: some columns, such as ‘out.electricity.total.energy_consumption’ and ‘out.site_energy.total.energy_consumption’, are summations of other columns. This is important to keep in mind as you plot the data.

What do you have to do?

With this understanding, you may begin the two-part assignment. Deliverables for each section are in blue, with the associated points in red, along with the associated points for their completion. Your submission should include the answers to each deliverable, along with the Python code that helped you complete this assignment. Please ensure you clearly reference the section you are answering by its number.

Should you need any guidance, do not hesitate to schedule a meeting with me: benvergara96@gwu.edu

Assignment Outline ***(100 points total)***

All the analysis should be done in Python, with the help of ChatGPT to avoid spending too much time reading code documentation.

1. **Data Processing:**
   1.1. **Identify Corrupted Data:** Each of the three building datasets contains 30 features (columns), of which the latter 25 are metered load data in kWh, representing various types of consumption. Each column comprises data points for 15-minute intervals throughout the year, totaling 35,040 points. Your task is to ensure that these 25 columns are free from corrupted or missing data.

   Fill out the following table: *(10 points)*

File Name Containing Corrupted Data	Row Number Containing Corrupted Data	What is the corrupted data?
.	.	.
.	.	.
.	.	.

   1.2. **Plan your Mitigation Strategy:** Think about how you should address this data point, based on the fact that these data points are ordered (time-series data) and reported in kWh.

   Answer the following questions: *(5 points each)*

   1.2.1. What would happen if you replaced it with a 0?
   1.2.2. What would happen if you deleted that row completely?
   1.2.3. What is the correct way to deal with this corrupted data?

   1.3. **Mitigate Errors through Linear Interpolation:** Fill in the erroneous data point with a linear interpolation between the preceding and superseding data points.

   Fill out the following table: *(15 points)*

File Name Containing Corrupted Data	Row Number Containing Corrupted Data	Preceding Data Point	Superseding Data Point	Replaced Number of Corrupted Data
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

2. **Data Visualization:**

2.1. **Plot Selection:** Spend some time on this website identifying three plots that you could use to represent any aspect of this data. You do not have to use all columns/data points. You can choose what you would like to focus on. For example, if you are curious about how much of the total electricity goes to fans, pumps, or lighting, a pie chart that represents this distribution would suffice as a plot type, even though you didn't use other columns, like natural gas consumed.

Fill out the following table: *(20 points)*

Plot Type	What aspect of the data does this plot type represent?	Why do you think this aspect is worth representing?
.	.	.
.	.	.
.	.	.

2.2. **Generate Plots:** Code the generation of your selected plots.

Attach three plots for each of the three datasets, totaling nine plots. *(20 points)*

2.3. **Plot Analysis:** Comment on what you observe in these plots.

Answer the following questions: *(10 points each)*

2.3.1. For each plot type, what is a unique conclusion or insight it offers that other plots do not?
2.3.2. Now that you have different plots for each of the three buildings, what can you notice about the differences in building energy use? What reasons could justify the differences in consumption?