Project 1:

Wyett Considine
Kal Parvanov
Brian Morales

Energy Consumption and Efficiency
https://www.eia.gov/consumption/commercial/about.php

Use the available data to determine energy usage trends across the states to track usage and efficiency.

- Directions to the data:
  - follow the link
  - Click the data tab at the top of the page under the primary header
  - Click the microdata tab in the secondary set of dropdowns
  - The data we are using is 2018 CBECS microdata
  - The explanation of the variables in the Variable and response codebook

Motivation

Why is this an important and interesting problem? What triggered the idea?

- Over the last few years in multiple states there have been numerous energy infrastructure issues in the US. With large swings in temperature and notable heat waves, along with the introduction of more electric vehicles, energy infrastructures and allocation of energy related resources have been put under a lot of stress. Because of this, blackouts and brownouts have become a more pressing issue.

A description of why it would be useful to use / develop machine learning to solve this problem.

- If someone could use machine learning to see if there are patterns in the kinds of energy usage, and the features of a building, they could create a list of modifications buildings could use to improve their energy consumption, and reduce the magnitude of power spikes, which are harmful to the energy infrastructure.

Data / Data Plan

A description of your data?

- We will be working with the U.S. Energy Information Administration public data set that contains untabulated records about energy consumption in individual buildings across the U.S. The data set includes information about the square footage of a building, electricity consumption, natural gas usage, if it is solar powered, number of floors and many more attributes. The table holds over 1000 features which we will reduce to a much smaller more focussed set of variables.

What are some of the interesting or critical features you have?

- A lot of the variables are binary, or trinary in nature, which implies the use of random forests or at least variations of decision trees. For example, there are a large number of variables that deal with whether the building uses a certain energy source for a certain task. There is a list of common energy sources, like propane, solar, or coal, and a list of energy uses, like heating, cooling, and equipment usage. Combinations of these variables provide key insights into how the building uses energy, and the resources they use the most of.
- In addition, there are a lot of variables detailing the amount of energy used, which will be critical in the kind of analysis we hope to use this data for.

Are there any features you plan to exclude? Approximately (or exactly) how many samples do you have?

Yes, since our dataset incorporates over 1000 features we plan to exclude more than a third of the dataset. We mainly want to focus on predictors that take into account electricity usage. Many of the features are imputation flags and of an unknown eligibility. We don’t have a strong background in what imputation flags and unknown eligibility are and therefore we will be excluding them.