**Capstone project for the Data Analysis Pathway: “Model development to predict total energy use of an office building.”**

**Joe Potts**

**Overview**

This project is designed to predict the electricity and natural gas consumption of an office building using a machine learning model generated from the data sources. Due to confidentiality concerns, the dataset has no location data for the surveyed buildings so detailed analysis of a city or state cannot be done. The whole state of Kentucky is in climate zone 4 along w/ several other states and sections of states located from the west coast to the east coast. Climate zone 4 is further divided into three humidity levels: 4A-mixed humidity, 4B-mixed dry, and 4C-mixed marine. These humidity indicators (A, B, or C) are not included in the CBECS survey. Kentucky is in climate zone 4A.

**Features**

* Selection of variables to be considered: climate zone, total energy consumption, building square footage, building age, and heat/cooling plant type are expected to be some of the main drivers of the energy consumed.
* Correlation of the variables to total energy cost.
* Summary of building features in climate zone 4.
* Data request from a public API.

**Technical details**

**Data sources**

US Energy Information Administration Commercial Buildings Energy Consumption Survey (CBECS) microdata file from the 2018 data collection (most recent available). <https://www.eia.gov/consumption/commercial/data/2018/index.php?view=microdata> This study is done every six years. The 2024 study is in development now. The 2018 microdata file is a valuable tool for users conducting detailed analysis of energy use in commercial buildings. It contains greater than 6,400 rows, about 620 survey variables and about 630 imputation flags and weighting variables.

<https://openweathermap.org/api>. This resource will be used to add more detailed weather characteristics such as min/max temperatures and average temperature for the survey year of 2018.

**Below are the tasks and tools used to complete the analysis.**

* Analysis tools will include python, sql, pandas, mathplotlib, sklearn.
* Analysis plan
* Load CBECS data into a DataFrame file.
* API request for the weather data needed.
* Load the weather data into an SQL table.
* Choose the needed data columns such as square footage, building age, etc. from the whole CBECS data frame and save this as the working file.
* Join the weather data table to the working file data.
* Clean the working file data.
* Further limit the number of variables by using only the data with the highest correlations to total energy used. Strive to get the fewest variables possible.

Tasks continued…

* Split the working file into training and testing datasets.
* Fit the model using the training dataset.
* Generate coefficients and intercept values and do manual testing.
* Run the test dataset w/ the model and evaluate.
* Make the model usable for a user.