1. Project Description

You have been hired by an energy company (eSC). eSC provides electricity (power) to residential properties in South Carolina (and a small part North Carolina).

eSC is concerned about global warming, specifically the impact of global warming increasing the demand for their electricity. In short, they are worried that next summer will put too much demand on their electrical grid (ex. their ability to supply electricity to their customers when they want to cool their homes). If this happens there will be blackouts, which eSC wants to avoid at all costs.

Their goal is to understand potential energy usage during an extra hot summer.

Since they are focused on the summer, eSC is focused on July energy usage. July was selected, as eSC thinks that July is typically the highest energy usage month.

Background that might be useful:

There are several factors that might impact the cooling of a house, such as:

1. Climate (temperature, humidity, wind speed, etc), which can significantly impact cooling needs of the house.
2. Orientation of the house, since direction of the house or the placement of windows, can impact the amount of sunlight and shade the house receives.
3. The amount and quality of insulation in the house can impact the effectiveness of the heating and cooling systems. Insulation in a house is done at each level such as ceiling, floor, walls, roof, and windows. In USA, the insulation in described by an R-value. The higher the R-Value, the greater the insulating power and resistance to heat flow. The insulation power varies for different materials. Compared to steel or masonry, wood is a better option for improving energy efficiency when used in wall assemblies due to its naturally higher thermal resistance.
4. The size of the house can impact the heating and cooling needs, as larger houses require more energy to heat and cool than smaller houses

2. The following data has been provided

2A. Static House Data

A file with basic house information for a random sample of single-family houses that eSC serves.

Specifically, this file contains the list of all 150 houses in the dataset. For each house, there is information describing the house. This information ranges from the building id (used to access the energy data mentioned below) to other house attributes that do not change (such as the size of the house).

Hint: there are lots of columns that are not interesting – you need to determine which columns you think are important

The file can be found on blackboard, and is in 'csv' format.

2B. Energy Usage Data

Energy usage data - for each house, energy usage data, which was collected hour-by-hour.

There is one dataset file per house.

The dataset consists of calibrated and validated energy usage, with 1 hour load profiles. In other words, within one file, the data describes the usage of energy from many different sources (ex. air conditioning system, dryer), per hour for that house.

Each file contains individual timeseries data of a specific house, with the 'building ID' as file name which identifies the house.

Note that each file is in 'parquet' (an optimized for storage CSV file) format.

All the data is in one folder on amazon AWS. For example, the following URL is for 'building_id' 8414.

https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/2023-houseData/8414.parquet

All 150 houses are in the same directory.

2C. Meta Data

A data description file, explaining the fields used across the different housing data files.

In other words, this is a simple, human readable, file that contains a description of the attributes (that are in either the static data or the energy usage data.

https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/data_dictionary.csv

2D. Weather Data

Hour-by-hour weather information (one file for each geographic area)

The timeseries weather data was collected for each county and stored based on a county code.

The county code for each house can be found at 'in.county' column of the house static dataset. This file is in a simple CSV format.

For example, the following URL provides the weather for county ‘G4500010’.

https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/weather/2023-weather-data/G4500010.csv

There are approximately 50 counties in the directory.

3. You have the following tasks

a) Determine the best approach to read and merge the data and determine what should be the output during this ‘data preparation’ phase. Some hints:

   a. You need to merge the weather data and the house energy usage data. This will create 150 merged files.
      i. You can use the ‘bldg_id’ to find the appropriate house data
      ii. You can use the ‘in.county’ attribute to find the appropriate weather data
      iii. Only focus on July. Since you only need to focus on July, make sure to only include / save data for that month.

   b. Create a new column – total electrical energy usage (by summing the relevant attributes into a new total electrical energy usage).

   c. One approach is to create one larger dataframe, based on the 150 merged data files (in other words, combine all the rows into 1 dataframe). Make sure each row has some of the key static house information for that house. Example important static information might include:
      i. Building id
      ii. in.sqft

Once you have a merged dataframe:

b) Do exploratory analysis of the data – to gain some basic insight about the data (focusing on what might impact total electrical energy usage)

c) Build a model that predicts the energy usage

   You can try several models and pick the best model.

d) Understand and be able to explain your model’s accuracy.

e) Create a new dataset, with all July temperatures 5 degrees warmer

f) Use your best model to evaluate peak future energy demand (assuming no new customers)
    a. Note: this must be model driven, not just increasing energy usage by a percentage

g) Create visualizations of future peak energy demand in total (for an hour):

4. Your need to deliver

a) The code of your analysis

b) A presentation (to the CEO of the power company) explaining the results of your analysis