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 on 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.

Rather than build out the capability to deliver more energy to their clients (i.e., build another power plant), they want to understand the key drivers of energy usage, and how they could encourage their customers to save energy.

In short, their goal is to reduce energy usage if next summer is 'extra hot', so that they can meet demand (and not build a new energy production facility). This approach would also help the environment!

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

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 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.

The file can be found at:

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

There are around 5,000 houses in the file.

Note that this file is in 'parquet' (an optimized for storage CSV file) 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’ 102063.

https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/2023-houseData/102063.parquet

There are approximately 5,000 houses (i.e., different building id’s) in the 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

1. Determine the best approach to read and merge the data and determine what should be the output during this ‘data preparation’ phase.

2. Do exploratory analysis of the data – to gain some basic insight about the data.

3. Build a model that predicts the energy usage, for a given hour, for the month of July. July was selected, as eSC thought July is typically the highest energy usage month.

   Hint: you will need to try several models and pick the best model.

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

5. Create a new weather dataset, with all July temperatures 5 degrees warmer.

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

7. Show future peak energy demand in total (for an hour):
    a. For different geographic regions
    b. For other dimensions /attributes you think important

8. Create a shiny application so that your client can interact with the data
    a. To better understand your model's energy prediction
    b. To better understand the potential future energy needs (and drivers of that future energy need)

9. Identify one potential approach to reduce peak energy demand.

10. What would you suggest, how would you model the impact? How would you explain the impact? BE DATA DRIVEN!

4. Your team needs to deliver

1. The **code** of your analysis. The code file can be .R, .Rmd, or .ipynb.

2. A **presentation** (to the CEO of the power company).

3. A URL for the **shiny app**.

4. A **document** explaining the work done (will be reviewed by your 'technical manager').

   → This document must also explain which team member did which tasks.

5. An update every two weeks (one per group, not per person). For each update (including for the final submission), provide:

   a. Work done by each person (since the last update).
   b. Work planned to be done by each person (by the next update).
   c. Key issues/challenges.

   → this should be the basis for your deliverable in item (4), the document explaining the work done (including for example, models that were bad, but still evaluated)