# **General Information**

## Role:

• You should act as a consultant

## Goal:

• How might your client predict future energy usage?

– Predict energy usage if the summer was 5 degrees warmer

* Energy usage = things that affect our 5-6 variables + temperature + sqft
* 4 ac, 4 heating + temp + sqft

– Provide actionable insight into how to reduce energy costs

Use our 5-6 principal outcome.

* ac: 4 variables
* heating: 4 variables
* - use decision tree

A) Determine the best approach to read and merge the data

B) Determine what should be the output during this ‘data preparation’ phase.

C) **Build a model that predicts the energy usage, for any 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.

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

## What data to analyze – should be:

• A function of what the team determines might be useful

• Determined by each project team

• There is \*A LOT\* of data

Remember this needs to be data driven

## Project Deliverables

**Word Document:**

• Target audience is your manager / instructor

(hint: your manager/instructor is a data science expert)

• Focus on what was accomplished

• Should describe all analysis done, even if an analysis did not generate any interesting results, it should still be included

**Presentation:**

• Target audience is your client *(hint: the client is not a data science expert)*

• Presentation length is 10 minutes *(lab instructor will explain specifics)*

• Be sure to include the following in your presentation:

Number of records in dataset evaluated

Key drivers identified; accuracy of results

**Interactive Application (shiny app):**

a) A shiny app needs to be created and deployed on shinyapps.io

b) To better understand your model’s energy prediction

c) To better understand the potential future energy needs and/or savings

## Project Updates

1) Project Updates: Nov 2, Nov 16, Nov 30 (one per group, not per person)

2) 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

# **Project Data**

## House Data

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

* The file contains the list of all houses in the dataset.
* For each house, there is information describing the house.
  + The information ranges from the building id (used to access the energy data) to other house attributes that do not change (such as the size of the house).
* There are ~5,000 houses in the dataset (rows in the file)
* The file can be found at:
  + <https://intro-datascience.s3.us-east-2.amazonaws.com/SCdata/static_house_info.parquet>

*Note that this file is in ‘parquet’ (an optimized for storage CSV file) format.*

## Energy Usage Data

For each house, there is a file that contains energy usage data, which was collected hour-by-hour.

* There is one data file per house. Energy usage is:
  + Collected every hour
  + Collected across many sources (ex. air conditioning system, dryer
  + the ‘building ID’ is file name which identifies the house.
* Note that each file is in ‘parquet’ (an optimized for storage CSV file).
* 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*

## Weather Data

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

• The 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’.

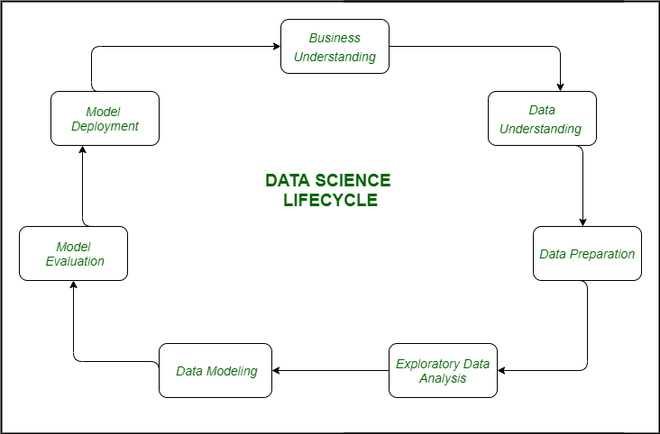
<https://intro-datascience.s3.us-east-2.amazonaws.com/SCdata/weather/2023-weather-data/G4500010.csv>

## Metadata

* A data description file, explaining the fields used across the different housing data files.
* ~270 attributes
* <https://intro-datascience.s3.us-east-2.amazonaws.com/SCdata/data_dictionary.csv>

# Data Science Process

* <https://www.geeksforgeeks.org/data-science-lifecycle/>
  + The globally mentioned structure in fixing any analytical problem is referred to as a Cross Industry Standard Process for Data Mining or CRISP-DM framework.



**1. Business Understanding:** The complete cycle revolves around the enterprise goal. What will you resolve if you do not longer have a specific problem? It is extraordinarily essential to apprehend the commercial enterprise goal sincerely due to the fact that will be your ultimate aim of the analysis. After desirable perception only we can set the precise aim of evaluation that is in sync with the enterprise objective. You need to understand if the customer desires to minimize savings loss, or if they prefer to predict the rate of a commodity, etc.

**2. Data Understanding:** After enterprise understanding, the subsequent step is data understanding. This includes a series of all the reachable data. Here you need to intently work with the commercial enterprise group as they are certainly conscious of what information is present, what facts should be used for this commercial enterprise problem, and different information. This step includes describing the data, their structure, their relevance, their records type. Explore the information using graphical plots. Basically, extracting any data that you can get about the information through simply exploring the data.

**3. Preparation of Data:** Next comes the data preparation stage. This consists of steps like choosing the applicable data, integrating the data by means of merging the data sets, cleaning it, treating the lacking values through either eliminating them or imputing them, treating inaccurate data through eliminating them, additionally test for outliers the use of box plots and cope with them. Constructing new data, derive new elements from present ones. Format the data into the preferred structure, eliminate undesirable columns and features. Data preparation is the most time-consuming but arguably the most essential step in the complete existence cycle. Your model will be as accurate as your data.

**4. Exploratory Data Analysis:** This step includes getting some concept about the answer and elements affecting it, earlier than constructing the real model. Distribution of data inside distinctive variables of a character is explored graphically the usage of bar-graphs, Relations between distinct aspects are captured via graphical representations like scatter plots and warmth maps. Many data visualization strategies are considerably used to discover each and every characteristic individually and by means of combining them with different features.

**5. Data Modeling:** Data modeling is the coronary heart of data analysis. A model takes the organized data as input and gives the preferred output. This step consists of selecting the suitable kind of model, whether the problem is a classification problem, or a regression problem or a clustering problem. After deciding on the model family, amongst the number of algorithms amongst that family, we need to cautiously pick out the algorithms to put into effect and enforce them. We need to tune the hyperparameters of every model to obtain the preferred performance. We additionally need to make positive there is the right stability between overall performance and generalizability. We do no longer desire the model to study the data and operate poorly on new data.

**6. Model Evaluation:** Here the model is evaluated for checking if it is geared up to be deployed. The model is examined on an unseen data, evaluated on a cautiously thought out set of assessment metrics. We additionally need to make positive that the model conforms to reality. If we do not acquire a quality end result in the evaluation, we have to re-iterate the complete modelling procedure until the preferred stage of metrics is achieved. Any data science solution, a machine learning model, simply like a human, must evolve, must be capable to enhance itself with new data, adapt to a new evaluation metric. We can construct more than one model for a certain phenomenon, however, a lot of them may additionally be imperfect. The model assessment helps us select and construct an ideal model.

**7. Model Deployment:** The model after a rigorous assessment is at the end deployed in the preferred structure and channel. This is the last step in the data science life cycle. Each step in the data science life cycle defined above must be laboured upon carefully. If any step is performed improperly, and hence, have an effect on the subsequent step and the complete effort goes to waste. For example, if data is no longer accumulated properly, you’ll lose records and you will no longer be constructing an ideal model. If information is not cleaned properly, the model will no longer work. If the model is not evaluated properly, it will fail in the actual world. Right from Business perception to model deployment, every step has to be given appropriate attention, time, and effort.