**Proposal**

**ENERGY CONSUMPTION PREDICTIVE MODEL**  
**Joseph Choi | Jenny Overby | Daniel Meier | Serge Nane**  
**DSC450 Applied Data Science  
Spring 2024**

# Domain

This project falls under the Energy Consumption and Sustainability domain, as our focus is to develop a predictive model that forecasts energy consumption based on historical data. Here are 10 references that we have compiled to help us make sense of the domain and the data we’ll be working with.

1. [Research on household energy demand patterns, data acquisition and influencing factors: A review - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S2210670723005279#:~:text=Different%20economic%20levels%2C%20household%20types%2C%20building%20structures%2C%20and,all%20exert%20an%20influence%20on%20household%20energy%20usage.): Highlights the importance of accurate data for energy efficiency and conservation and the main influencing factors.
2. [Determinants and approaches of household energy consumption: A review - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S235248472301168X): Looks into the effects of physical and social factors on household energy use and identifies key drivers and predictive models for energy forecasting.
3. [Household energy-saving behavior, its consumption, and life satisfaction in 37 countries | Scientific Reports (nature.com)](https://www.nature.com/articles/s41598-023-28368-8): Provides information on energy-saving behaviors and household income on energy consumption across 37 countries
4. [Frontiers | Research on energy consumption in household sector: a comprehensive review based on bibliometric analysis (frontiersin.org)](https://www.frontiersin.org/articles/10.3389/fenrg.2023.1209290/full): Conveys the importance of smart home technology and renewable energy sources for future energy-saving initiatives
5. [(PDF) Factors influencing electricity consumption: a review of research methods (researchgate.net)](https://www.researchgate.net/publication/328562123_Factors_influencing_electricity_consumption_a_review_of_research_methods): Analyzed factors influencing household electricity consumption
6. [Residential electricity consumption behavior: Influencing factors, related theories and intervention strategies - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S1364032117311164): Highlights the behaviors behind residential electricity consumption
7. [Household energy consumption, energy efficiency, and household income–Evidence from China - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S0306261923014381): Examines the relationship between energy consumption and household income in China
8. [Sustainability | Free Full-Text | Factors Influencing Energy Consumption in the Context of Sustainable Development (mdpi.com)](https://www.mdpi.com/2071-1050/11/15/4147): Explores factors influencing energy consumption within EU28 from 1995 to 2014
9. [Determinants of household energy choice for domestic chores: Evidence from the Semien Mountains National Park and Adjacent Districts, Northwest Ethiopia - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S2772783123000134): Examines household energy consumption levels, behaviors, and factors in Ethiopia
10. [Frontiers | Modeling Energy Consumption Using Machine Learning (frontiersin.org)](https://www.frontiersin.org/articles/10.3389/fmtec.2022.855208/full): Explores how the authors used machine learning to analyze over 15,000 samples in the U.S. Department of Energy. Identified Random Forest Regressor and Classifier as best-fit models for predicting energy consumption.

# About the Dataset

## Description

The dataset for this project is a list of time-stamped records that provide variables that influence energy usage in a building at a specific date and time. It is structured for conducting detailed analysis and modeling to identify trends, patterns, and correlations in energy consumption.

## Data Fields

* **Timestamp**: Date and time when the data was recorded
* **Temperature**: Degree of Celsius at the time of the recording
* **Humidity**: Amount of moisture in the air as a percentage
* **Square Footage**: Total area of the space, measured in square feet
* **Occupancy**: Number of people present in the space
* **HVAC Usage**: Indicates whether an HVAC system was used or not (On/Off)
* **Lighting Usage**: Indicates whether lighting systems were used or not (On/Off)
* **Renewable Energy**: Contribution of renewable energy to the total energy usage in percentage
* **Day Of Week**: Days ranging from Monday to Sunday
* **Holiday**: Indicates whether the data was recorded on a holiday or not (Yes/No)
* **Energy Consumption**: Measure that indicates total energy consumption

## Dataset Location

The dataset can be located and downloaded in the link: [Energy-consumption-prediction (kaggle.com)](https://www.kaggle.com/datasets/mrsimple07/energy-consumption-prediction?resource=download)

# Research Questions

Energy sustainability have long been a crucial topic of discussion as it directly impacts the health of our planet and the quality of life for current and future generations. To tackle this issue, our project approach is to address the following research questions:

1. What are the key indicators or factors influencing energy consumption?
2. Can we accurately predict future energy consumption in buildings using historical usage data?

Analyzing this data and getting insights into this business problem can lead to more efficient energy use, reduced costs, and support sustainability initiatives. Getting a good grasp of the factors influencing energy consumption empowers everyday energy users to make informed decisions and manage energy more effectively.

# Method

The plan for our project is to start with ensuring the quality and structure of our data. These initial steps include:

* **Data Cleaning**: Identifying and cleaning any inconsistencies, missing values, or outliers that may skew our analysis and model performance
* **Data Transformation**: Preprocessing raw data to be ready for analysis. The tasks involve scaling numerical features, encoding categorical variables, and handling date-time formats
* **Exploratory Data Analysis**: Conducting EDA to understand the dataset’s structure and identifying variables with high predictive power for energy consumption
* **Feature Engineering**: Improving the model's predictive capability by performing feature engineering on selected features

To address our research questions, our strategy is to analyze historical data and leverage machine learning to provide insights into energy efficiency and sustainability. This process includes:

* **Data Visualization**: Identifying trends, patterns, and correlations to display key indicators of energy consumption
* **Data Modeling**: Exploring and building different algorithms and evaluating their predictive performance

# Potential Issues

We found the dataset from Kaggle. Like many Kaggle datasets, it is synthetic, which means the data is designed to simulate real-world scenarios. Because the data is generated to emulate a hypothetical environment, there may be potential challenges with its accuracy.

Specifically, the “EnergyConsumption” feature, which also happens to be our target variable, is calculated using a simple formula by the author, but the exact calculation method is not disclosed. Therefore, this introduces some ambiguity regarding the accuracy of the values present in our target variable column. For our project, we will make assumptions that the “EnergyConsumption” values are calculated accurately based on the input variables and their correlations.

# Conclusion

Tying everything together, our project proposal is this. We would like to address our stated business problems head-on as a team, applying everything we’ve learned from our previous data science courses. Through carefully planned data cleaning, transformation, exploratory data analysis, feature engineering, and machine learning algorithms, we plan to thoroughly analyze key factors influencing energy consumption and develop predictive models that accurately forecast energy usage based on historical data. Despite the potential challenges of using synthetic data from Kaggle, we remain committed to ensuring data quality and making informed insights to drive our analysis forward. In doing so, we would like to gain a better understanding of efficient energy management and share our results.