1. Data Exploration and Preprocessing

- The data.csv file included 29 columns and 16,800 entries.
- Converted timestampto datetime format.
- The preprocessing step converted numerous numeric object-type columns to float format.
- Current data contains dropped rows where target values (`equipment_energy_consumption`) are missing.
- We substituted all other missing data points by applying column mean calculations.
- Extracted new features: hour day_of_week

2. Exploratory Data Analysis (EDA)

- A correlation heatmap depicting all numerical features helped in the analysis.
- The strongest correlations with the target variable were found to exist with 'lighting_energy' and temperature and humidity readings measured across multiple zones.
- `lighting_energy`
- Temperatures/humidity from several zones
- `outdoor_temperature`, `dew_point`

3. Feature Engineering and Selection

- Final feature matrix includes:
- The final feature set includes energy metrics alongside temperature readings and humidity rates and weather measurements.
- Extracted time-based features
- `random_variable1`, `random_variable2` evaluated for importanceThe team removed unimportant features from the matrix when necessary (see further information below).

4. Model Used

- Random Forest Regressor
- serves alongside
- Linear Regression as the principal models.
- A 80/20 Train-Test split provided the evaluation framework which utilized
- MAE, RMSE, R^2 metrics.
- 80/20 Train-Test Split Evaluated using: Using an 80/20 Train-Test Split we evaluated the models through MAE, RMSE and R^2.

5. Model Evaluation

```
THE R2 SCORE OF RANDOM FOREST REGRESSION MODEL 0.035876157706906
THE MEAN ABSOLUTE ERROR OF RANDOM FOREST MODEL 72.76870743005556
THE MEAN SQUARE ERROR OF RANDOM FOREST MODEL 3017.694835393137


THE R2 SCORE OF LINEAR REGRESSION MODEL 0.010804262721638369
THE MEAN ABSOLUTE ERROR OF LINEAR REGRESSION MODEL 75.2594687843362
THE MEAN SQUARE ERROR OF LINEAR REGRESSION MODEL 3091.681088822075
```

6.Feature Importance (Top 5)

1. Lighting Energy
2. Zone1 Temperature
3. Outdoor Temperature
4. Dew Point
5. Zone5 Temperature

The random variables `random_variable1` and `random_variable2` secured minimal rankings which caused them to be omitted from selection.

7. Recommendations

The energy consumption in high-impact zones (Zone 1 and Zone 5) should receive optimization measures.
Realign building light schedules according to times when usage peaks occurs.
The system needs to receive weather information through an outdoor temperature and dew point parameter for HVAC control adjustments.
Exclude non-informative features(`random_variable1/2`) from final model.

8. Summary

- The development of a strong baseline Random Forest model served as a starting point for our analysis.
- The feature importances analysis uncovered strategic factors which could lead to direct improvements.
- The predictive model needs operational context from machine load and shift patterns to achieve better performance.
- The system shows readiness for inclusion into energy management systems' operational pipelines.
-