# Report

## I. Context and challenges

Energy consumption is a major concern for companies because it directly impacts operating costs, environmental footprint, and regulatory compliance. High energy use leads to increased expenses, especially in industries that rely heavily on machinery, data processing, or transportation. Additionally, companies are under growing pressure to reduce carbon emissions and meet sustainability goals, both from governments through regulations and from consumers who prefer environmentally responsible brands

Thus, efficient energy use can also improve operational performance by identifying and eliminating waste. Overall, managing energy consumption is essential for cost savings, environmental responsibility, and maintaining a positive reputation.

We aim to use this project to propose concrete actions based on the results of the analyses, as optimization levers for managing energy consumption.

## II. Methodological approach

### a. Data collection

For this project, we chose a dataset on energy consumption, on Kaggle, specifically electricity usage data from 881 companies and local authorities across six French overseas regions: Réunion Island, French Guiana, Martinique, Guadeloupe, Mayotte, and Corsica. The data was collected between 2021 and 2024.

### b. Data preparation

Initially, our dataset contained 42,288 records and 112 columns. However, a significant number of columns had missing values. To address this, we removed all columns with more than 45% missing data. After this step, some columns still contained missing values, but the proportion was less than 2%. For these, we chose to impute the missing values using the column mean.

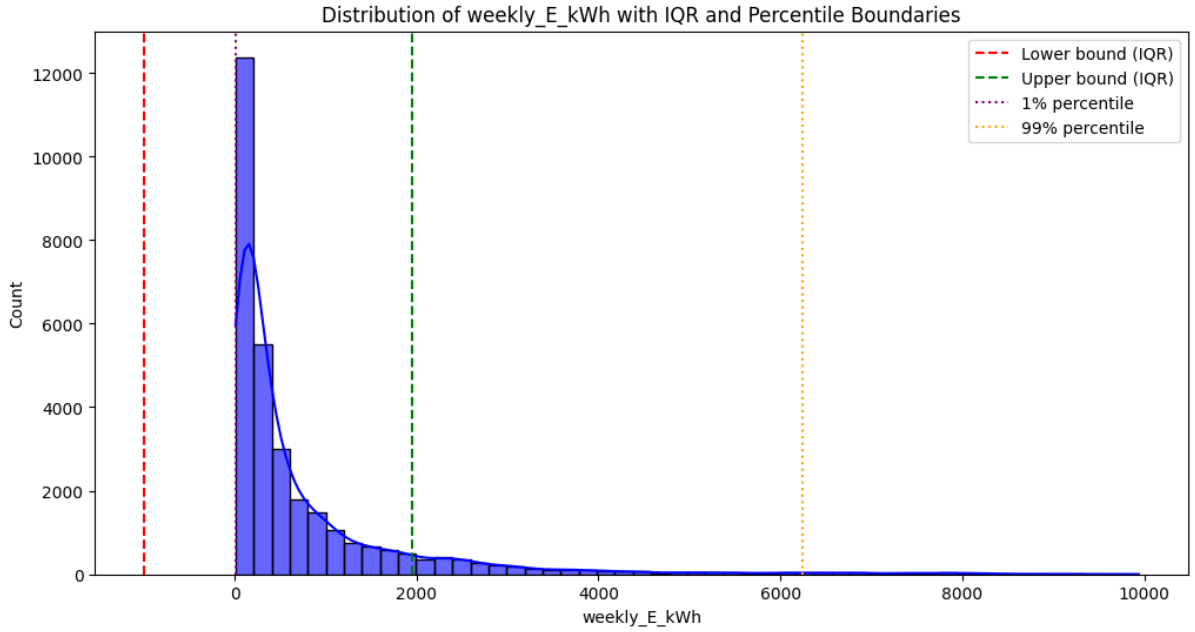
Since the feature weekly_E_kWh is the target column for prediction, we will delete all rows with missing values in this column. As a result, we ended up with 36,930 records and 22 columns.

Brief description of each column in our dataset:

- year-Wweek – The year and week number (e.g., 2024-W10) indicating the time period of the data.
- user_id – Unique identifier for each user.
- site_id – Unique identifier for each site (a building or a group of buildings).
- department – Name of the department where the sensor is located (e.g., South Corsica, Upper Corsica, Guadeloupe, etc.).
- nace_code – NACE code representing the type of activity of the company.
- insee_code – INSEE code representing the municipality.

- **weekly_E_kWh** – Weekly electricity consumption in kilowatt-hours (kWh).
- weekly_dd_heating_15 – Weekly Heating Degree-Days (HDD) based on the reference temperature 15°C, indicating heating demand.
- weekly_dd_heating_16 – HDD with a base temperature of 16°C.
- weekly_dd_heating_17 – HDD with a base temperature of 17°C.
- weekly_dd_heating_18 – HDD with a base temperature of 18°C.
- weekly_dd_cooling_22 – Weekly Cooling Degree-Days (CDD) based on the reference temperature 22°C, indicating cooling demand.
- weekly_dd_cooling_23 – CDD with a base temperature of 23°C.
- weekly_dd_cooling_24 – CDD with a base temperature of 24°C.
- weekly_dd_cooling_25 – CDD with a base temperature of 25°C.
- weekly_dd_cooling_26 – CDD with a base temperature of 26°C.
- mean_indoor_temperature_00 – The average indoor temperature for sensor 00 during the week.
- max_indoor_temperature_00 – The maximum recorded indoor temperature for sensor 00.
- min_indoor_temperature_00 – The minimum recorded indoor temperature for sensor 00.
- mean_indoor_humidity_00 – The average indoor humidity for sensor 00 during the week.
- max_indoor_humidity_00 – The maximum recorded indoor humidity for sensor 00.
- min_indoor_humidity_00 – The minimum recorded indoor humidity for sensor 00.

Next, we checked for the presence of outliers. The weekly_E_kWh column contained several outliers, which we chose to investigate further



The graph shows that weekly_E_kWh is highly right-skewed, with most values clustered near zero and a long tail extending toward higher values. The 99th percentile is significantly higher than the IQR upper bound, indicating extreme outliers. The lower bound appears negative, which is unrealistic for energy consumption and should be ignored. These outliers could impact modeling, and applying a log transformation or using robust models might help manage their influence.

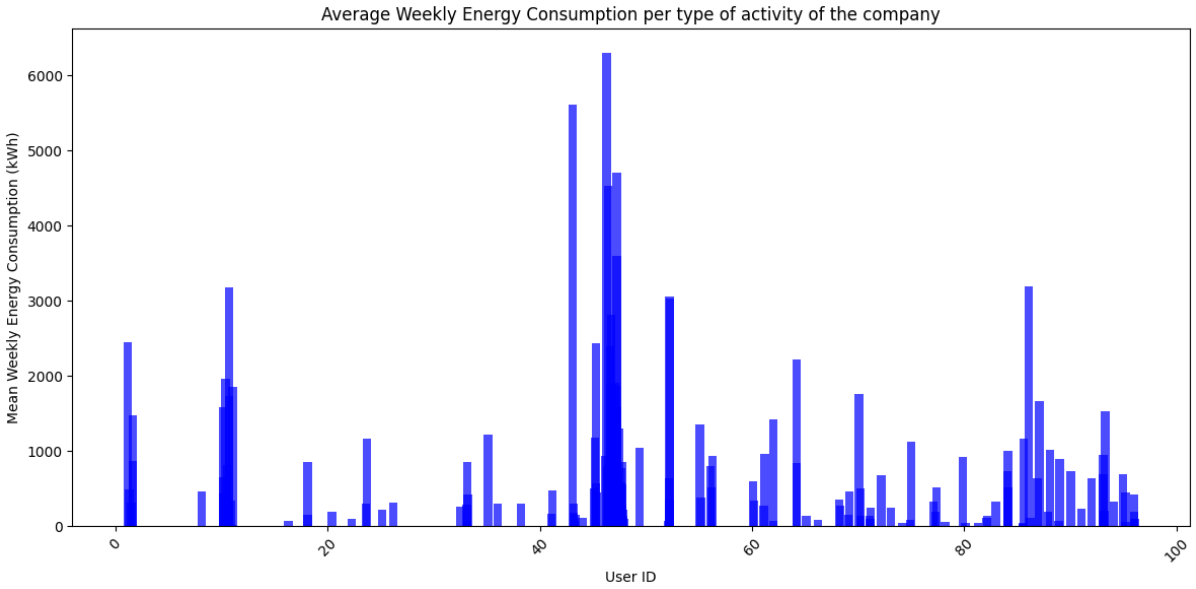
We decide not to do anything about this since some machine learning models like Decision Trees, Random Forests, and Gradient Boosting (XGBoost, LightGBM, CatBoost) handle outliers well.

To prepare the dataset for predictive and classification models, we encoded all non-integer features into numerical format.

### c. Analysis and Modeling

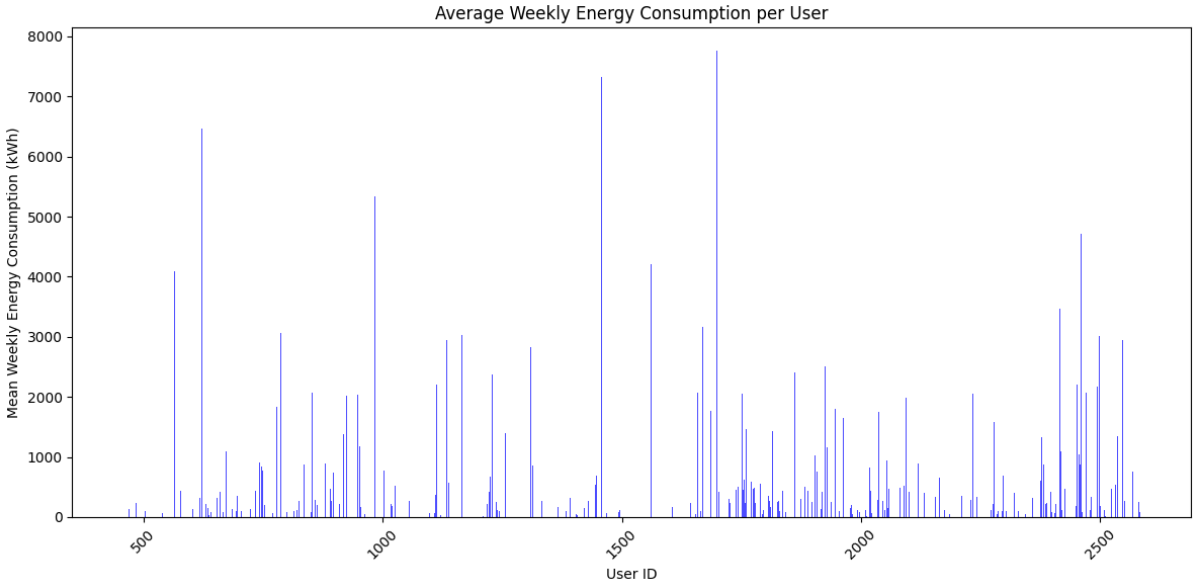
We visualized several graphs to gain a better understanding of the dataset and its underlying trends.

First, we visualized the average weekly energy consumption by company activity type. This graph helped us identify which activities were the most energy-intensive.



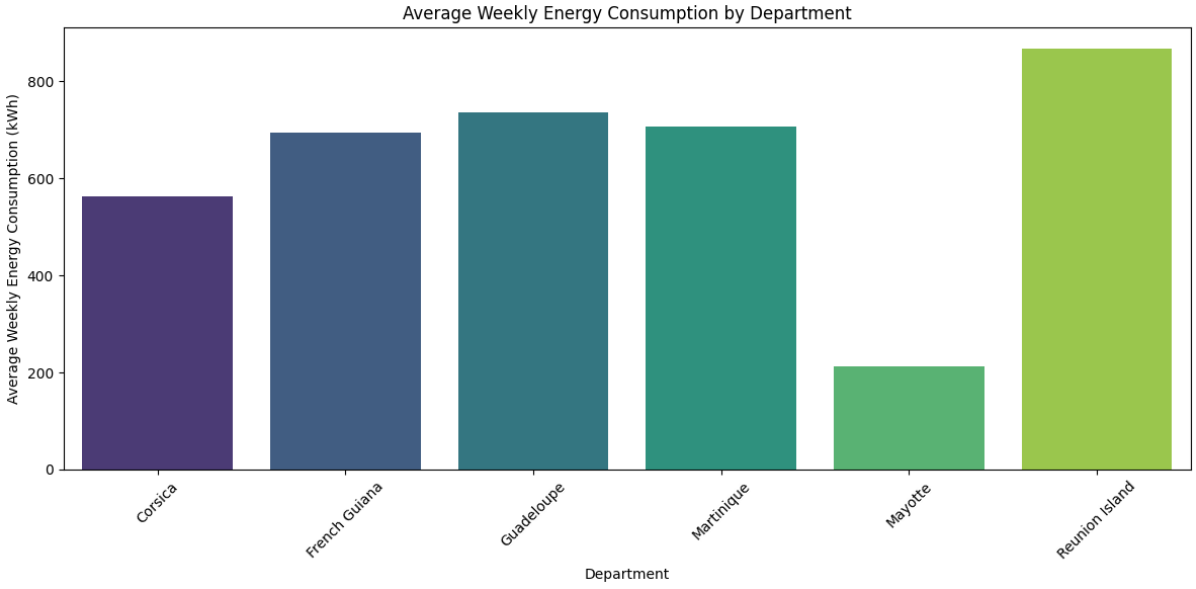

From this graph, we can conclude that energy consumption varies significantly across company activity types, with a few types showing notably higher average weekly energy use, indicating they are more energy-intensive. Most other activities consume relatively less energy in comparison.

Then, we visualized the average weekly energy consumption per user.




From this graph, we can conclude that most users have relatively low average weekly energy consumption, but there are a few users who consume significantly more energy, indicating potential outliers or high-demand users.

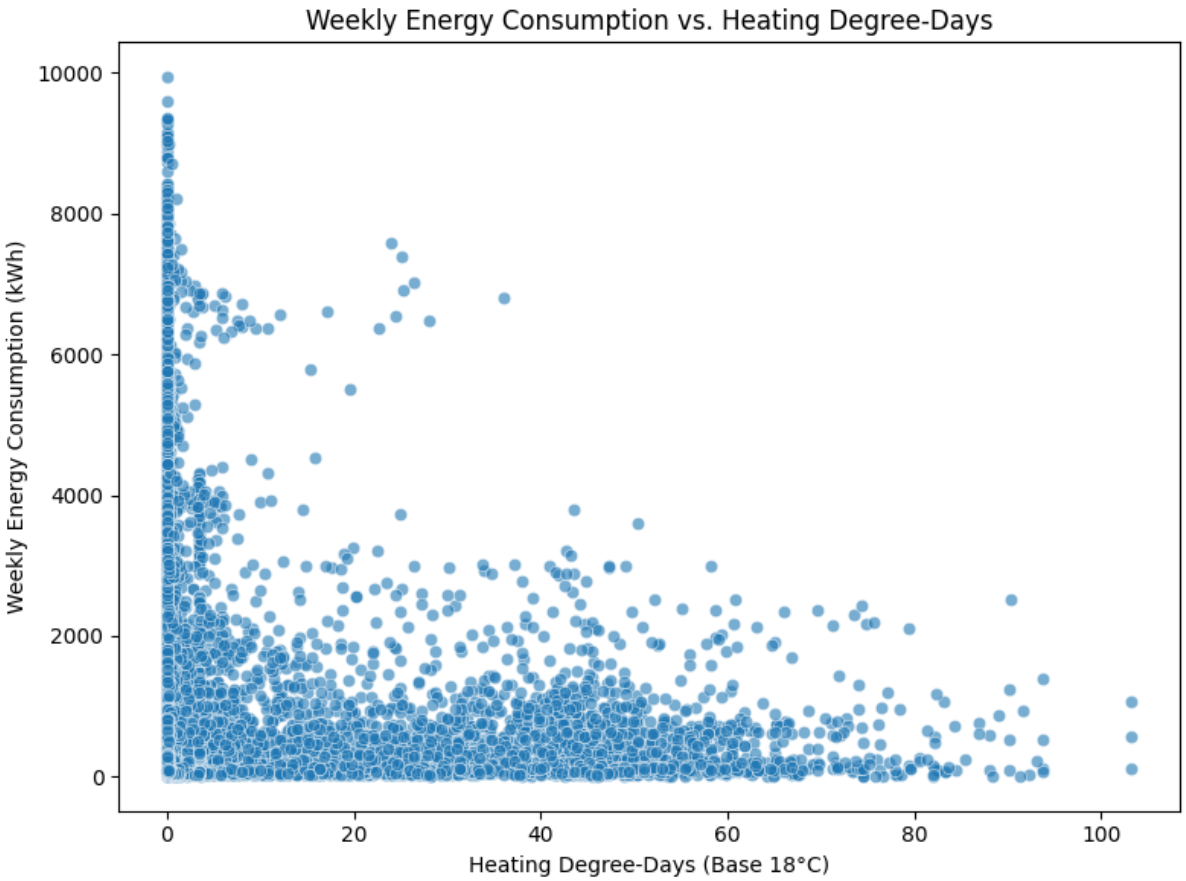
Next, we visualized the average weekly energy consumption by department to assess whether there were disparities in energy usage across different departments.




From this graph, we can conclude that there are clear disparities in average weekly energy consumption among departments. Notably, Réunion Island has the highest average consumption, while Mayotte consumes significantly less than the other departments. This suggests that geographic or operational differences may influence energy usage across regions.

We also generated a correlation heatmap to explore potential relationships between weekly electricity consumption and other variables in the dataset. However, no significant or noteworthy correlations were identified.

We then decided to compare the weekly energy consumption and the heating degree-days




It shows a weak negative trend, suggesting that energy consumption tends to decrease slightly as heating degree-days increase. However, the relationship is not strong or consistent, as there's a wide spread of consumption values across all levels of heating demand. Most notably, a large number of data points cluster around low heating degree-days, indicating that heating is not a dominant factor in overall energy consumption for many users.

### d. Predictive model for energy consumption

We decided to attempt predicting a company's weekly energy consumption based on the features available in our dataset. To do so, we experimented with several machine learning models, including Random Forest Regressor, XGBoost Regressor, LightGBM Regressor, Bayesian Ridge Regression, and a Deep Learning approach using Neural Networks.

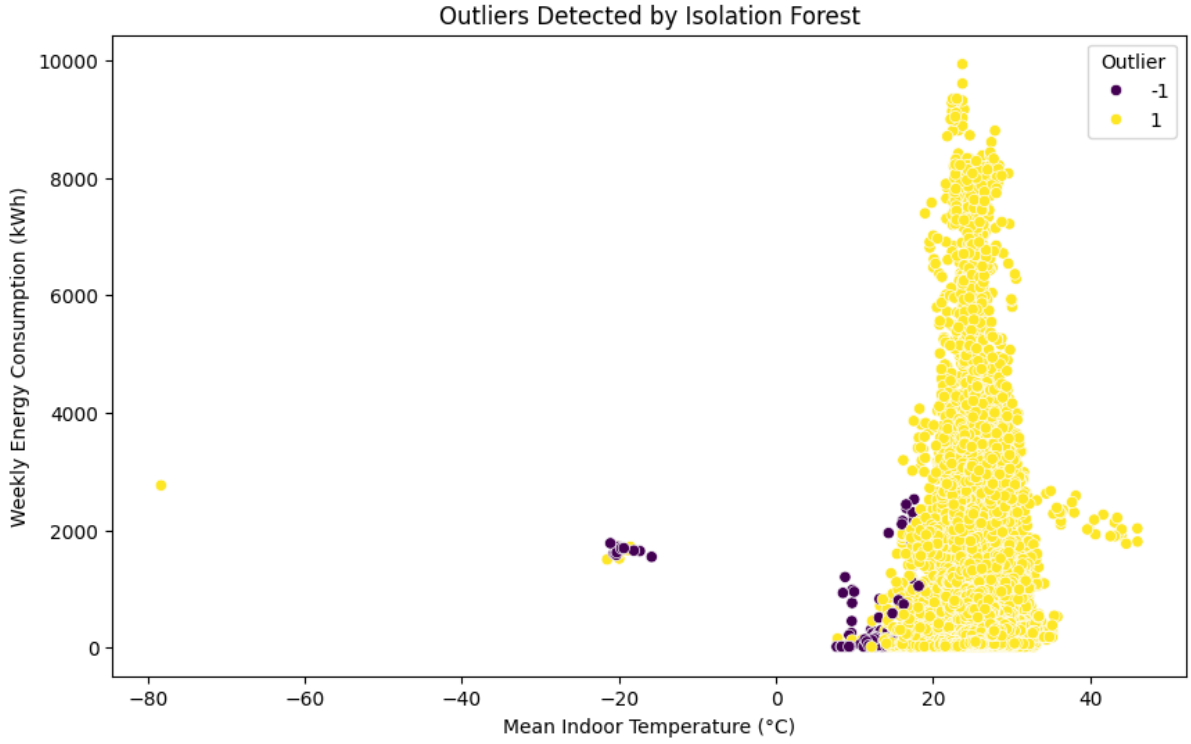
We experimented with several models, each chosen for its specific strengths. Random Forest Regressor was selected for its robustness, ability to handle non-linear relationships, and resistance to overfitting. XGBoost Regressor is well known for its high performance and efficiency, particularly with structured tabular data. We also included the LightGBM Regressor due to its optimization for speed and scalability, making it well-suited for large datasets with numerous features. Bayesian Ridge Regression was considered for its ability to capture uncertainty in predictions and its built-in regularization, which helps prevent overfitting. Finally, we explored Neural Networks, as they are capable of modeling complex, non-linear patterns—provided that a sufficient amount of data is available.

### e. Classification of High Consumption Periods

For the classification of high consumption periods, we created a new binary target variable: 1 for high consumption and 0 for low consumption. High consumption was defined as values falling within the top 25% of weekly energy consumption. To address the binary classification of high consumption periods, we explored a range of models, each chosen for its unique strengths. The Random Forest Classifier was selected for its strong performance, robustness, and ability to model complex, non-linear relationships without overfitting. We also used the XGBoost Classifier, known for its speed, scalability, and excellent results on structured data. The Support Vector Machine (SVM) Classifier was included due to its effectiveness in finding optimal decision boundaries, particularly in high-dimensional spaces. Additionally, we experimented with the K-Nearest Neighbors (KNN) Classifier, a simple and intuitive model that works well when similar instances share the same label. Lastly, the Naive Bayes Classifier was used as a fast and efficient baseline model, especially useful for handling categorical features and assuming feature independence.

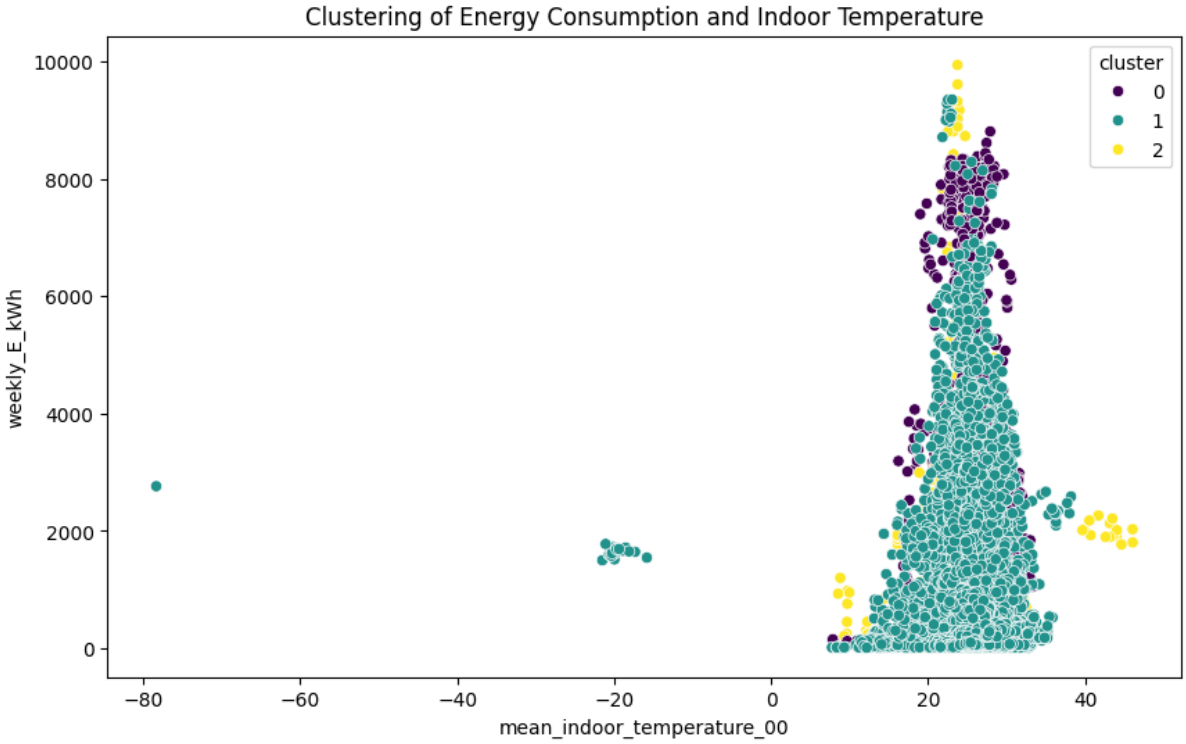
### f. Anomaly Detection for Energy Peaks

We attempted to detect anomalies in the data by identifying outliers and applying K-Means clustering for unsupervised anomaly detection. This was done to uncover unusual patterns or extreme consumption behaviors that could indicate data quality issues, abnormal usage, or potential opportunities for energy optimization.




This scatter plot shows the results of outlier detection using the Isolation Forest algorithm. Most data points (in yellow) are considered normal, while the algorithm identified a cluster of anomalous points (in purple), mainly located at the extremes of the mean indoor temperature axis, especially below 0°C and around 20°C with unusually low or high energy consumption. These outliers likely represent abnormal temperature readings or unusual consumption

behavior, and may indicate data entry errors or special operational conditions worth further investigation.




This K-Means clustering plot reveals three distinct groups based on energy consumption and indoor temperature. Most data falls into a central cluster, while smaller clusters highlight high consumption patterns and anomalous temperature values, potentially indicating unusual behaviors.

The Isolation Forest and K-Means clustering methods both helped identify unusual patterns in the data. Isolation Forest highlighted potential outliers, especially at extreme temperature values or unexpected consumption levels. Together, these two methods provided valuable insights for anomaly detection and a better understanding of energy usage patterns.

## III. **Presentation of results and insights**

### a. Regression model : Random Forest Regressor

Among all the predictive models tested for energy consumption, the Random Forest Regressor delivered the best results. It achieved an $R^2$ score of 0.957, indicating that it explains a large portion of the variance in the target variable. The model also produced reasonable error metrics, with a Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) that reflect solid predictive performance. On average, the model's residual error is approximately 237 kWh, suggesting good accuracy overall, though there is still some room for improvement.

To further validate the effectiveness of the Random Forest model, we applied cross-validation, which confirmed its strong performance. The regression model achieved a mean cross-validated RMSE of approximately 1,285 kWh, indicating a moderate prediction error and reinforcing the model’s reliability and generalization capability.

### b. Classification model : XGBoost Classifier

Among all the classifiers tested for identifying high consumption periods, the XGBoost Classifier achieved the best performance. It reached an overall accuracy of 98%, with a precision of 0.97 and recall of 0.96 for the high consumption class (label 1), resulting in a strong F1-score of 0.97. These metrics indicate that the model is highly effective at correctly identifying both high and low consumption instances, with minimal false positives or false negatives. Overall, XGBoost demonstrated excellent balance between precision and recall, making it the most reliable choice for this classification task.

To enhance the performance of the XGBoost Classifier, we applied SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance. The resulting SMOTE + XGBoost Classifier achieved near-perfect performance, with precision, recall, and F1-score all at 0.99 for both classes, indicating exceptional balance and accuracy in classifying energy consumption patterns. Additionally, we performed cross-validation on the Random Forest Classifier, which confirmed its robustness, achieving a mean cross-validated accuracy of 98.1%, reflecting strong generalization and near-perfect class discrimination.

## IV. Recommendations

### a. Target High-Consumption Activities

The analysis identified specific types of activities that are significantly more energy-intensive.So , we should try focusing on energy audits and efficiency improvements in these high-demand sectors (e.g., better insulation, energy-efficient machinery, or usage scheduling during off-peak hours).

### b. Regional Prioritization

Departments like Réunion Island have significantly higher energy consumption than others. We recommand to develop targeted regional energy-saving programs, such as renewable energy incentives, or awareness campaigns in the most consuming areas.

### c. Predictive Energy Monitoring

The Random Forest Regressor performed well in predicting weekly consumption. We recommand to integrate this model into an energy monitoring system to forecast consumption in real time and identify unusual usage before it becomes costly.

### d. Automated Anomaly Detection

Isolation Forest and K-Means models effectively detected unusual patterns and outliers. We could use these models in production to flag anomalies, such as sudden spikes or abnormal temperature readings, enabling early intervention or maintenance checks.

### e. Encourage Energy Awareness

In a strategy of awareness, we could share model insights with operational teams to raise awareness about high-consumption behaviors.