CASE STUDY :

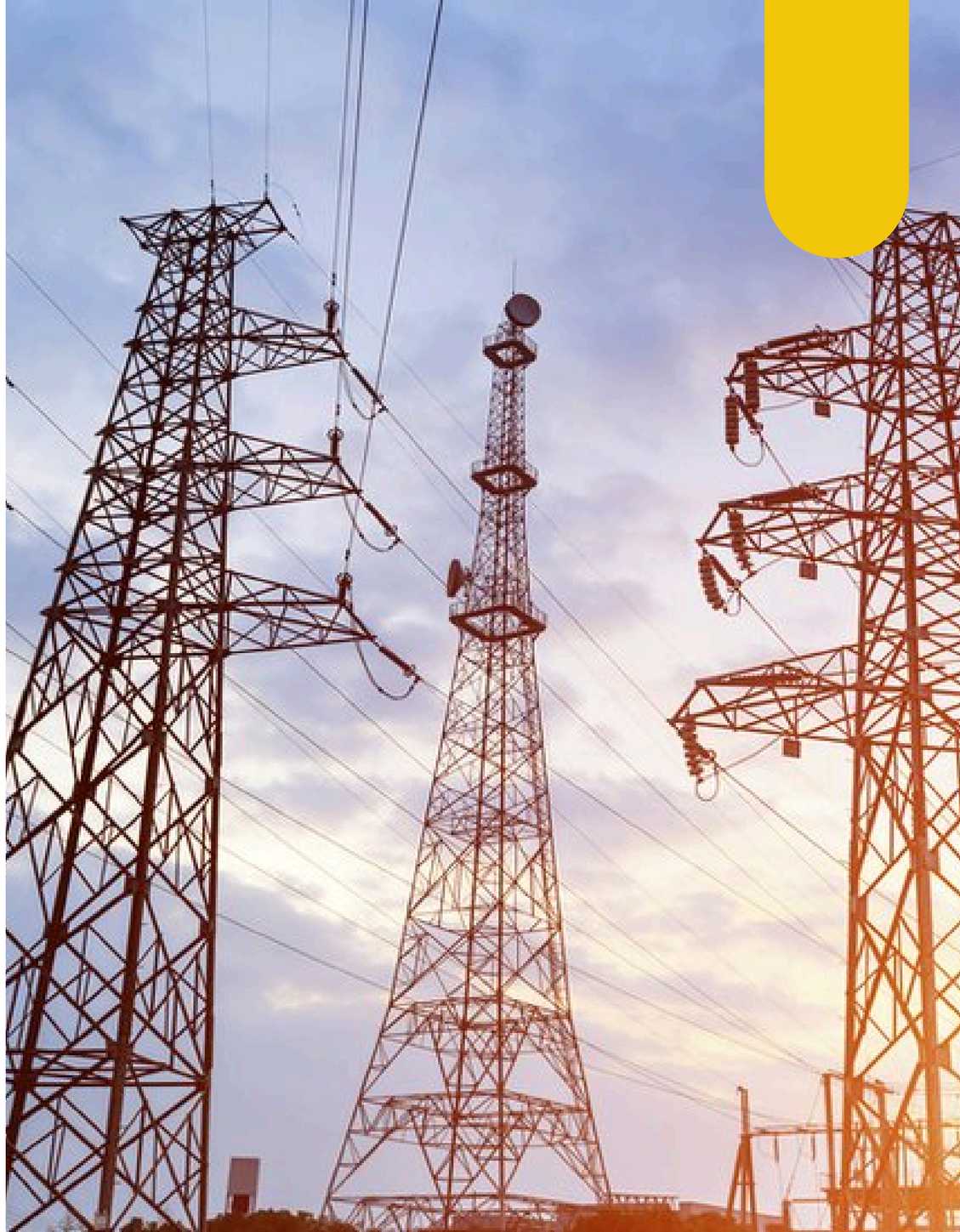
ANALYSIS AND OPTIMIZATION OF ENERGY CONSUMPTION

- NGOULAYE KEUNGUEU (DIA)
- ISMAEL KONE (DIA)
- DJIBRIL LALEG (DIA)
- LISA NACCACHE (DIA)
- LEINA PRIEUR (DIA)
- KYLIE WU (OCC)
- ILAN ZINI (DIA)

CONTEXT

In the context of the energy transition and global warming, businesses and organizations must minimize their environmental footprint while making efficient use of their resources. Managing energy consumption is a key strategic lever in addressing these global challenges.

Goal: use a dataset to predict the weekly energy consumption of a company.



PHASE 1: DATA COLLECTION

Data collection: our dataset on energy consumption, 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.

```
data.shape
```

```
(42288, 112)
```

Our dataset contains 42,288 records and 112 columns

PHASE 1: DATA PREPARATION

Missing values : We decide to delete all columns that have more than 45 % of their values missing. We deleted 90 columns.
We still have a very small pourcentage of missing values for some of our remaining features so we decide to replace the missing values in those features by the mean value of the corresponding feature.

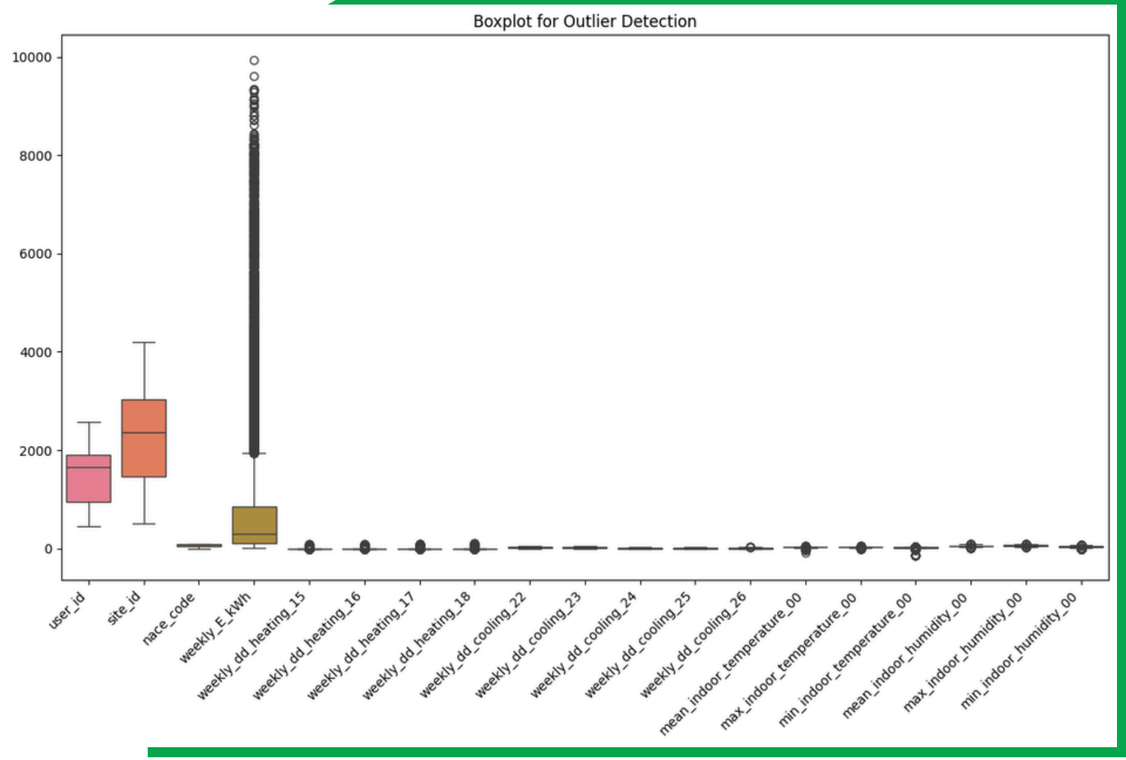
After cleaning, our dataset contains **30,725 rows** and **22 columns**.

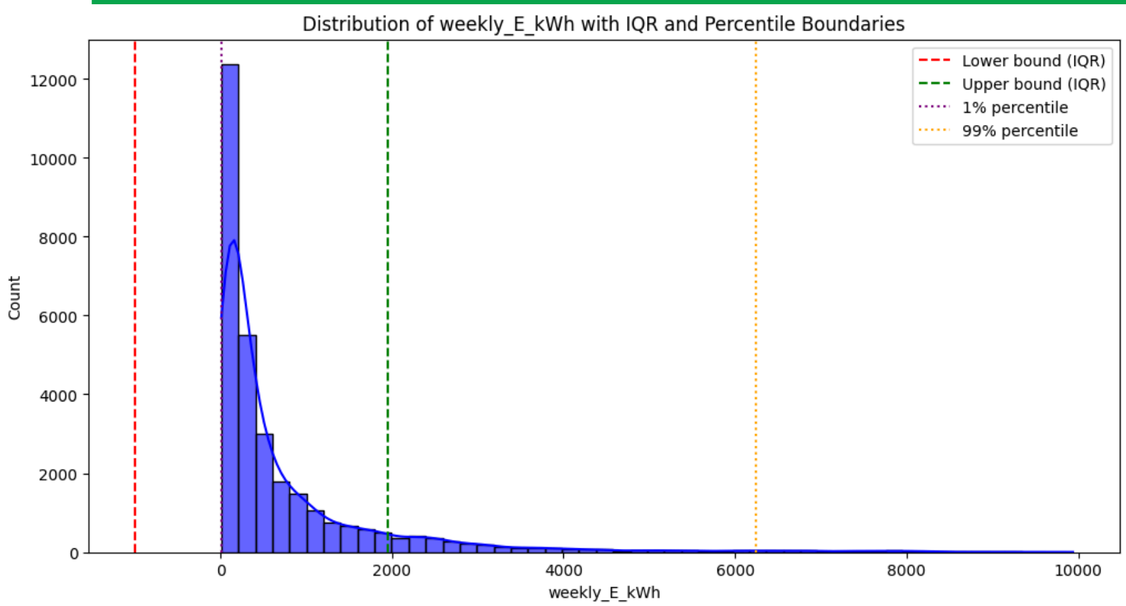
PHASE 1: DATA PREPARATION

Outliers management : We can see that the target column, Weekly Electricity Consumption, contains outliers.

We plotted of the Distribution of Weekly electricity consumption with IQR and Percentile Boundaries.

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.





PHASE 1: DATA PREPARATION

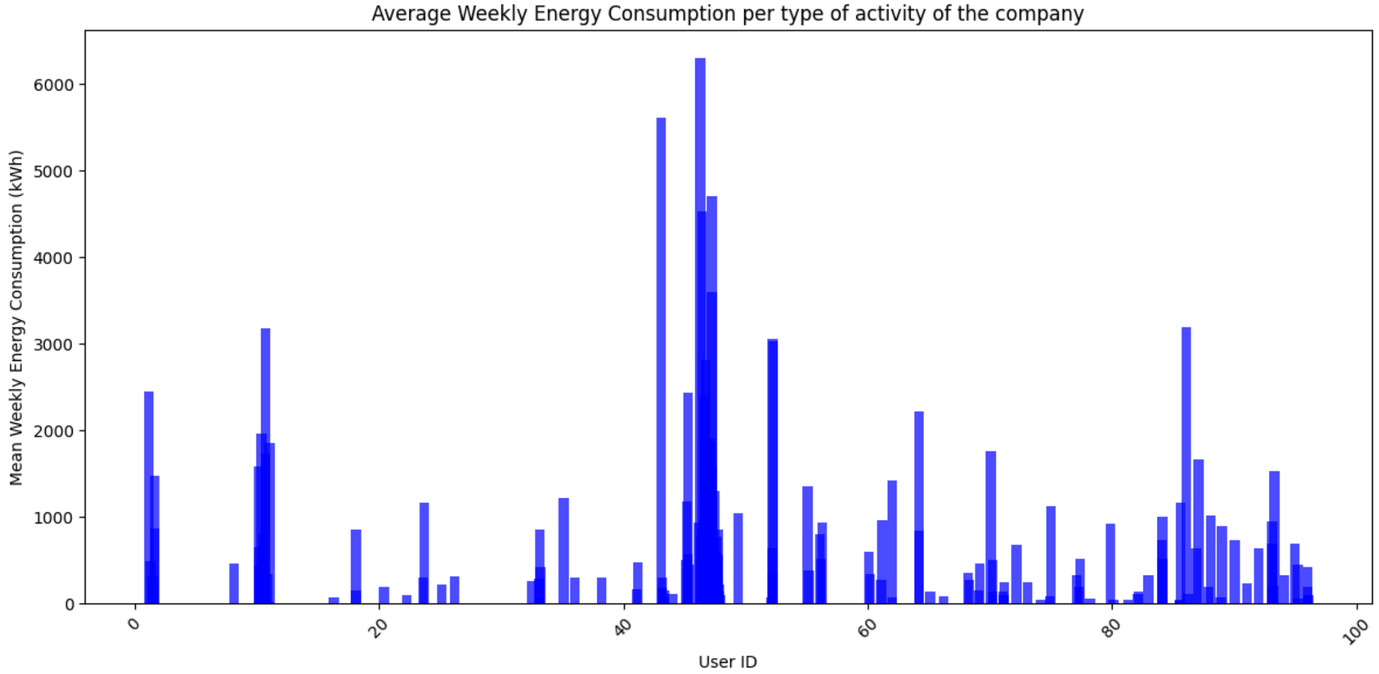
Feature Engineering / Encoding :

PHASE 1: DATA PREPARATION

Missing values : We decide to delete all columns that have more than 45 % of their values missing. We deleted 90 columns.
We still have a very small pourcentage of missing values for some of our remaining features so we decide to replace the missing values in those features by the mean value of the corresponding feature.

After cleaning, our dataset contains **30,725 rows** and **22 columns**.

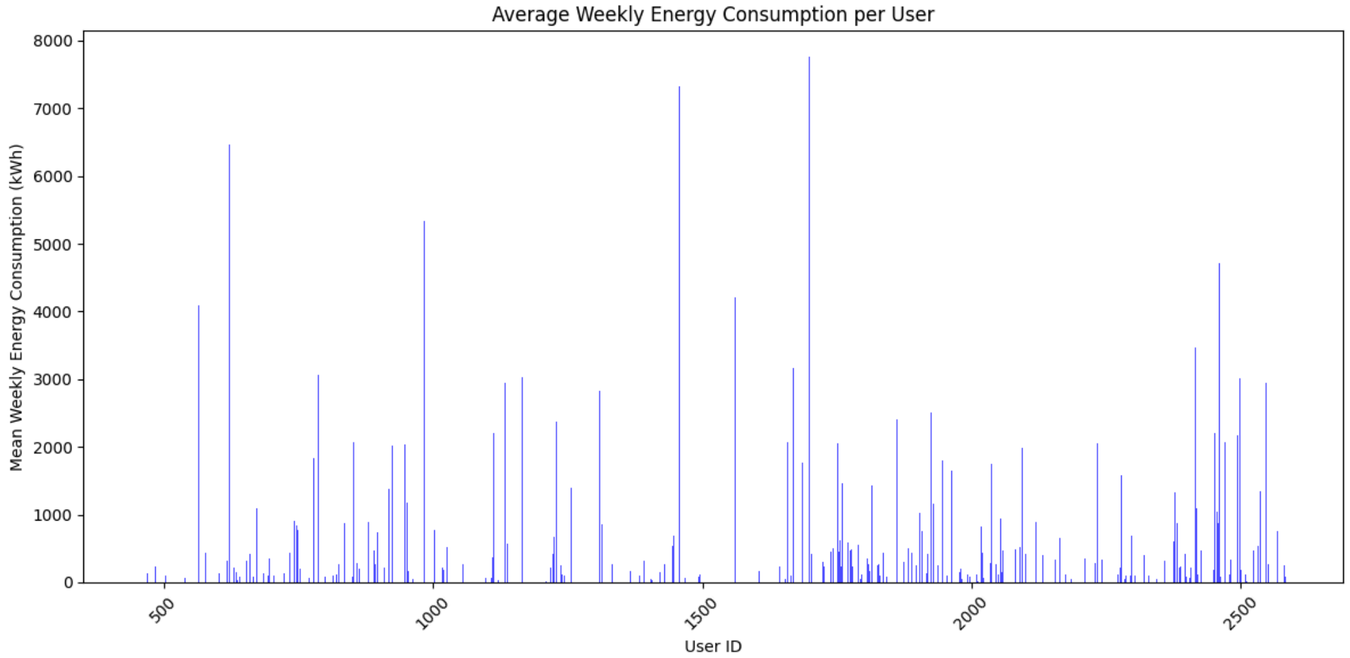
PHASE 2: Analysis and Modeling



Exploratory data analysis (EDA):
Average weekly energy consumption per type of activity of the company

Observation : 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.

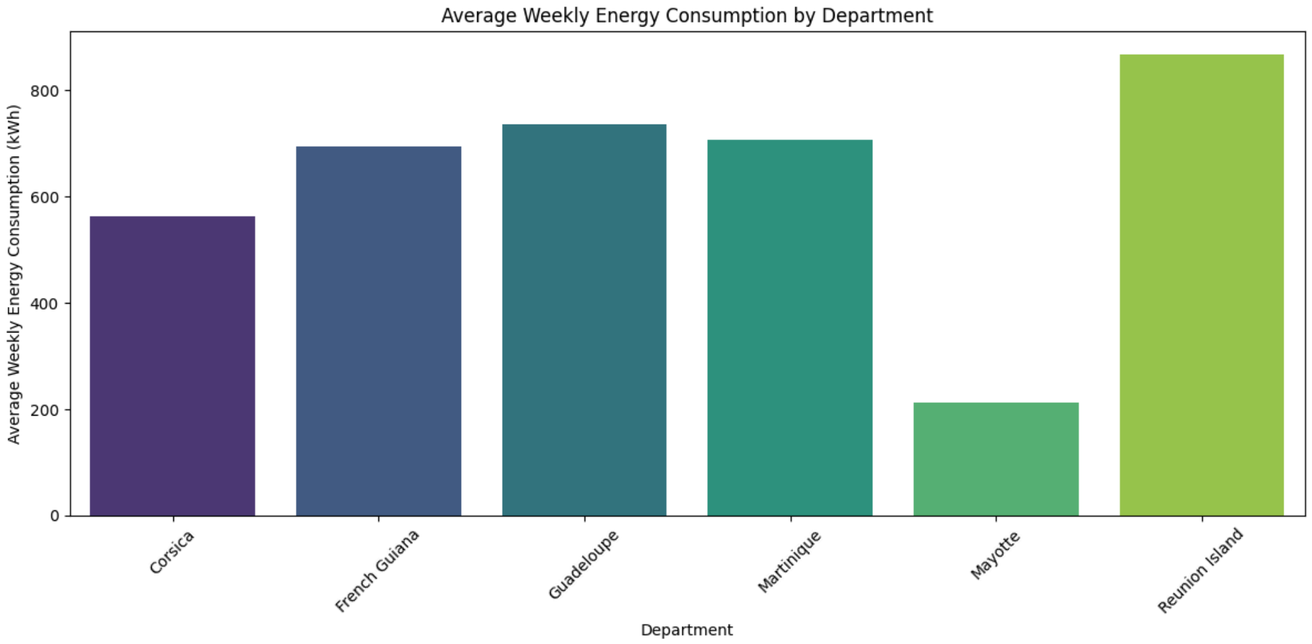
PHASE 2: Analysis and Modeling



Exploratory data analysis (EDA):
Average Weekly Energy Consumption per User

Observation: Most users have relatively low average weekly energy consumption, but there are a few users who consume significantly more energy, indicating high-demand users.

PHASE 2: Analysis and Modeling



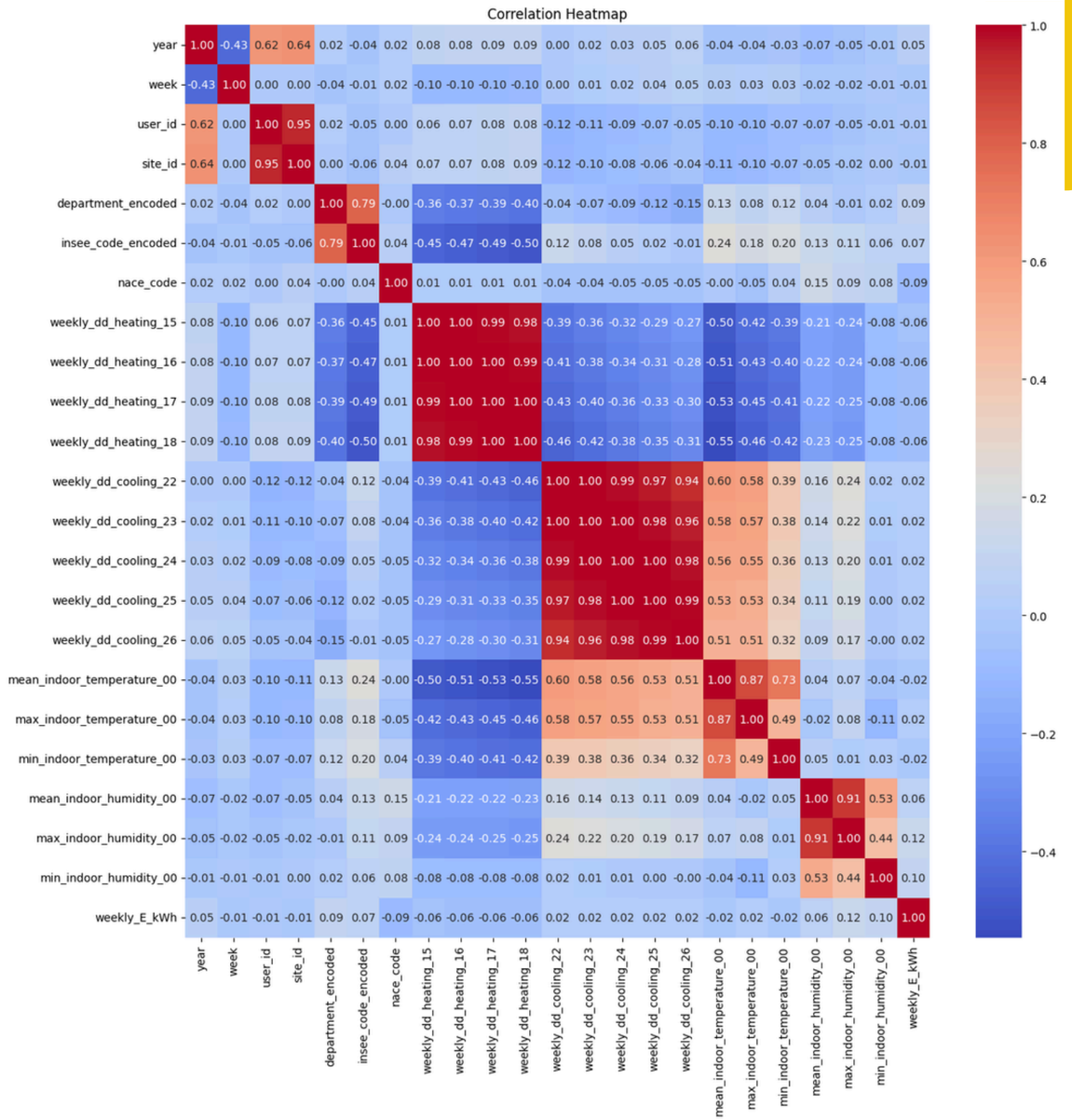
Exploratory data analysis (EDA):
Energy consumption by department

Observation: Clear disparities in average weekly energy consumption among departments. Notably, Réunion Island has the highest average consumption. This suggests that geographic or operational differences may influence energy usage across regions.

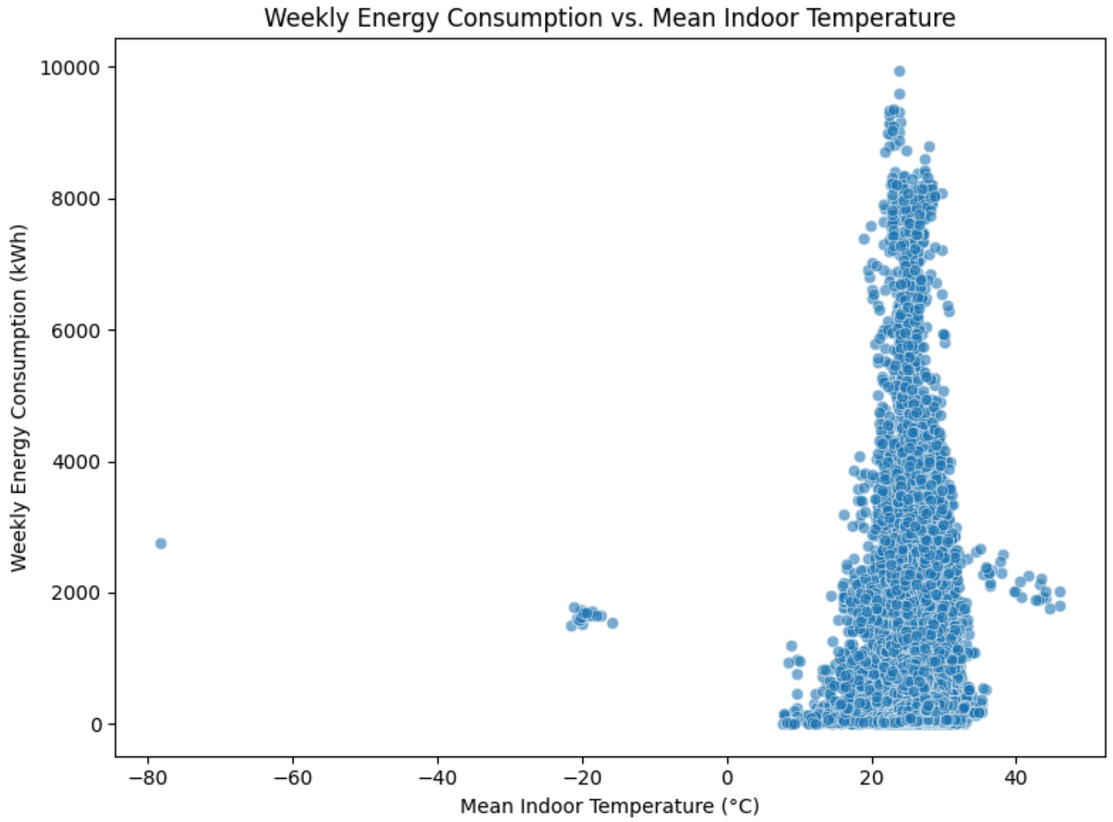
PHASE 2: Analysis and Modeling

Exploratory data analysis (EDA):
Correlation Heatmap

Observation: No significant or noteworthy correlations identified.



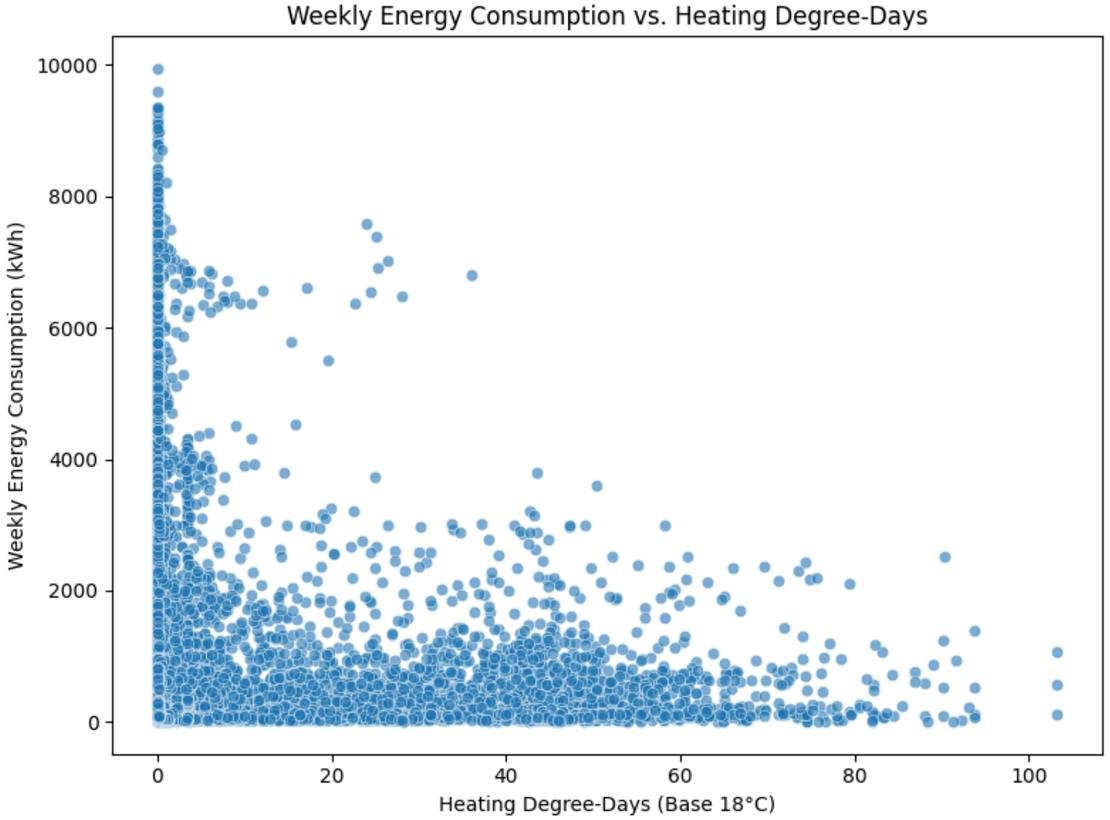
PHASE 2: Analysis and Modeling



Exploratory data analysis (EDA):
Weekly Energy Consumption vs. Mean Indoor Temperature

Observation: This scatter plot shows a strong concentration of data between 10°C and 30°C of mean indoor temperature, with most weekly energy consumption values below 4000 kWh.

PHASE 2: Analysis and Modeling



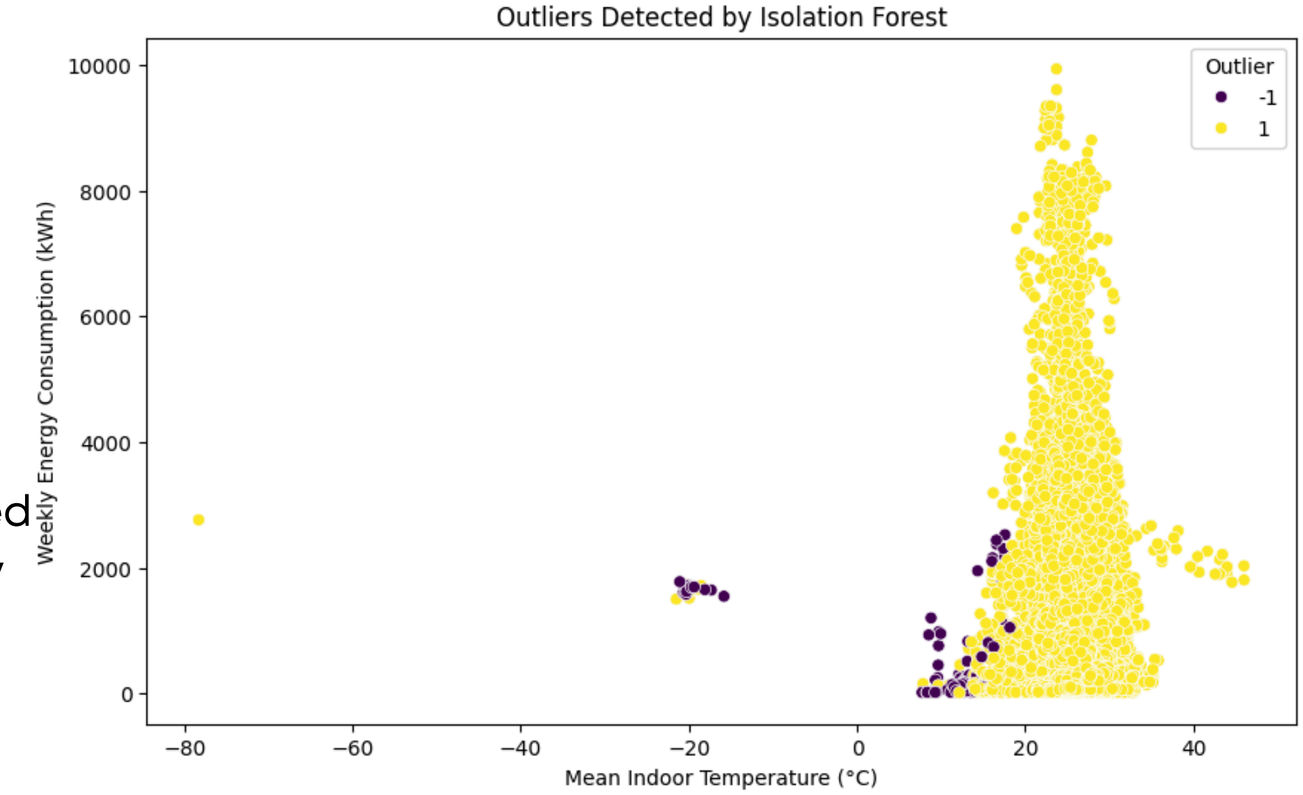
Exploratory data analysis (EDA):
Weekly Energy Consumption vs. Heating Degree-Days

Observation: Shows a weak negative trend, suggesting that energy consumption tends to decrease slightly as heating degree-days increase. 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

PHASE 2: Analysis and Modeling

Anomaly Detection for Energy Peaks:

Outliers Detected by Isolation Forest

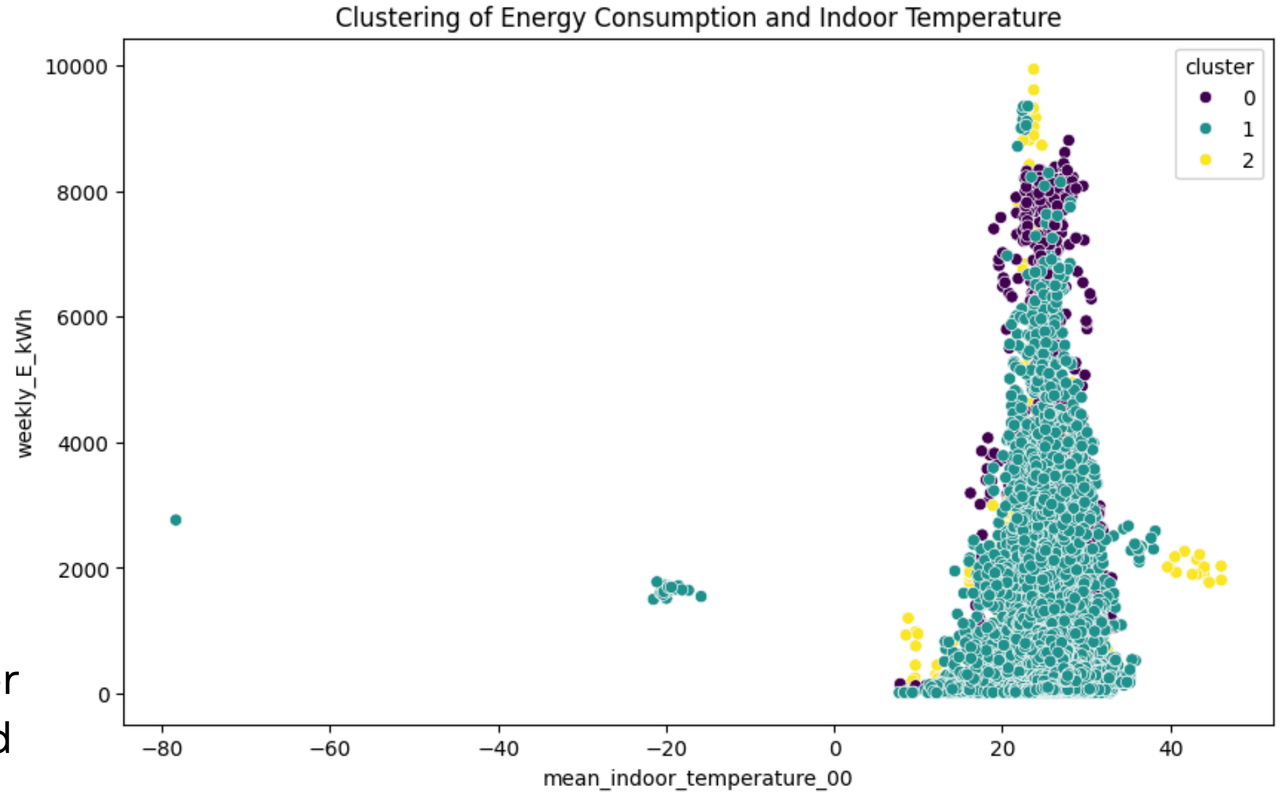


Observation: Most data points (in yellow) are considered normal, while the algorithm identified a cluster of anomalous points (in purple), mainly located at the extremes of the mean indoor temperature axis, especially below 0°C and around 20°C with unusually low or high energy consumption. These outliers likely represent abnormal temperature readings or unusual consumption behavior, and may indicate data entry errors or special operational conditions worth further investigation

PHASE 2: Analysis and Modeling

Anomaly Detection for Energy Peaks:

Clustering of Energy Consumption and Indoor Temperature using K-Means Clustering



Observation: 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.

PHASE 2: Analysis and Modeling

Predictive model for energy consumption:

Model	MAE	RMS	R2	REASON
Random Forest Regression	100.9	56534.3	0.96	Robustness, ability to handle non-linear relationships, and resistance to overfitting
XGBoost Regressor	334.4	546.9	0.77	High performance and efficiency, particularly with structured tabular data
LightGBM Regressor	188.6	348.2	0.91	Optimization for speed and scalability, making it well-suited for large datasets with numerous features
Bayesian Ridge Regression	690.2	1112.2	0.06	Ability to capture uncertainty in predictions and its built-in regularization
Neural Networks	598.7	1246.5	-0.17	Modeling complex, non-linear patterns

PHASE 2: Analysis and Modeling

Anomaly Detection for Energy Peaks:

Model	Accuracy	F1-score(0)	F1-score(1)	REASON
Random Forest	0.97	0.98	0.93	Strong performance, robustness, and ability to model complex, non-linear relationships without overfitting
XGB Classifier	0.98	0.99	0.97	Speed, scalability, and excellent results on structured data
SVM Classifier	0.75	0.86	0.0	Effectiveness in finding optimal decision boundaries, particularly in high-dimensional spaces.
KNeighbors Classifier	0.98	0.98	0.95	A simple and intuitive model that works well when similar instances share the same label
Naive Bayes Classifier	0.63	0.73	0.39	Useful for handling categorical features and assuming feature independence.

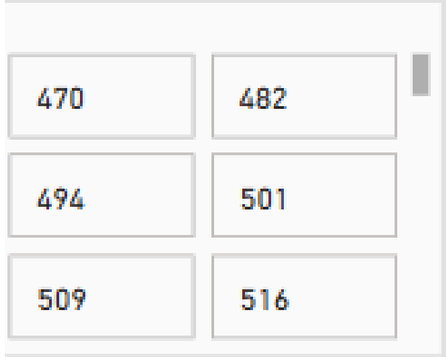
PHASE 2: Analysis and Modeling

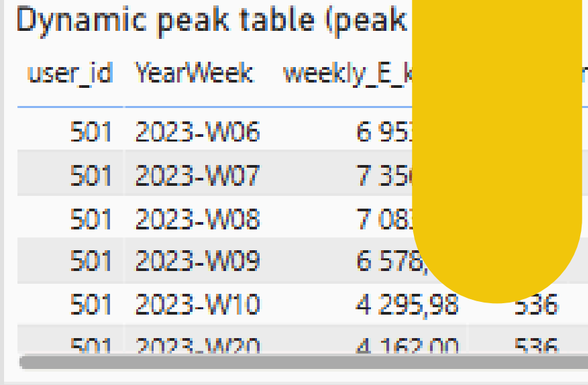
Cross-validation :

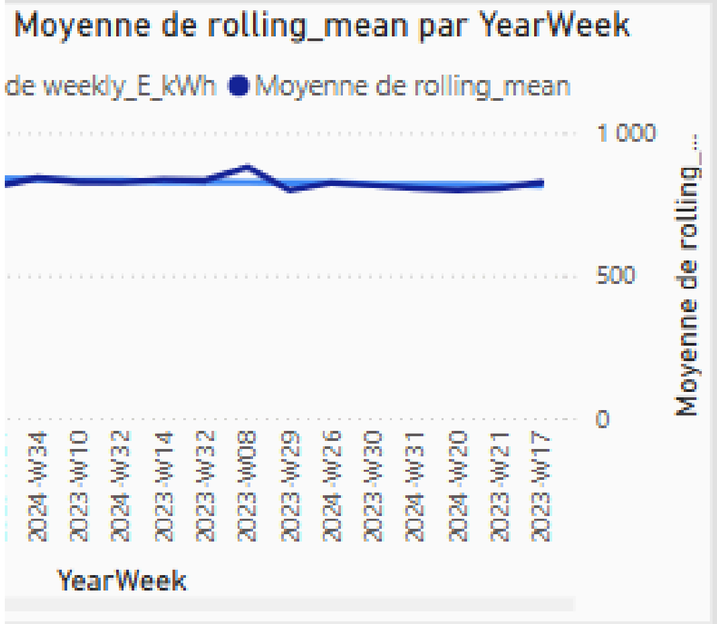
	Cross-validation	Improvement
Random Forest Regression	Mean Cross-Validation RMSE: 1287.6	
XGB Classifier	Mean Cross-Validation Accuracy: 0.97	Class imbalance : SMOTE + XGBoost Classifier -> 0.99 precision, recall, and F1-score on both classes

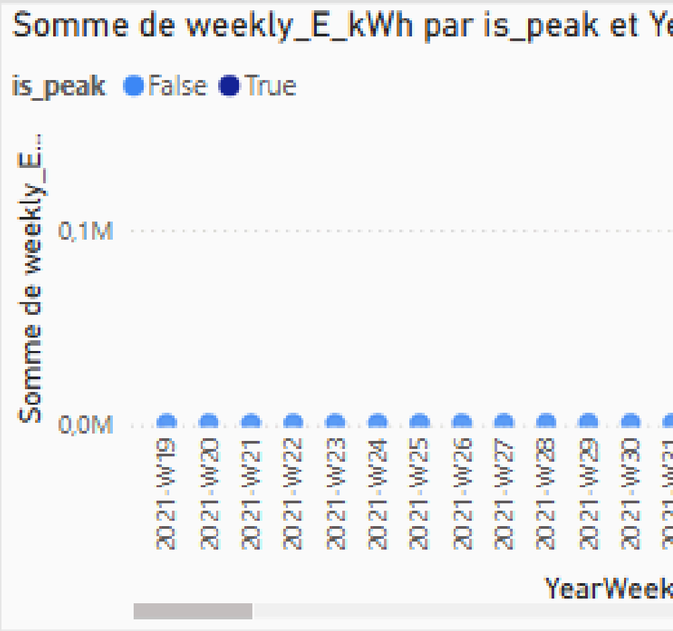
PHASE 3: CREATING A DASHBOARD

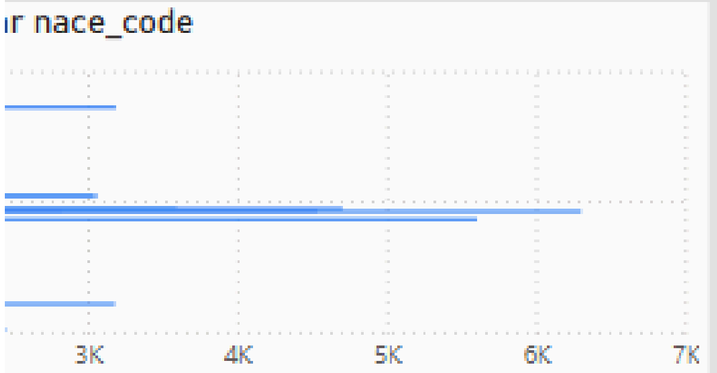
To visualize and analyze energy consumption trends, peaks, and patterns, we used Power BI to create an interactive dashboard based on the cleaned and processed dataset.

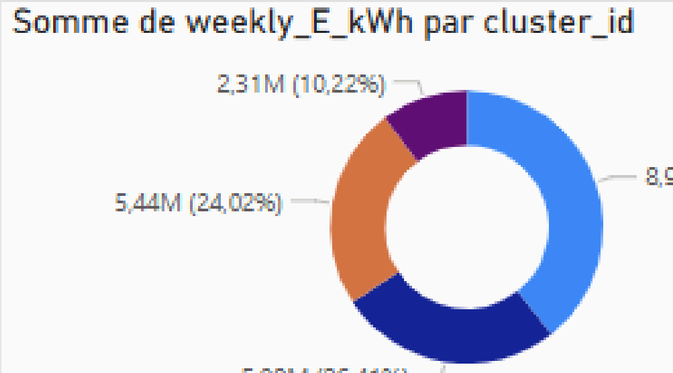












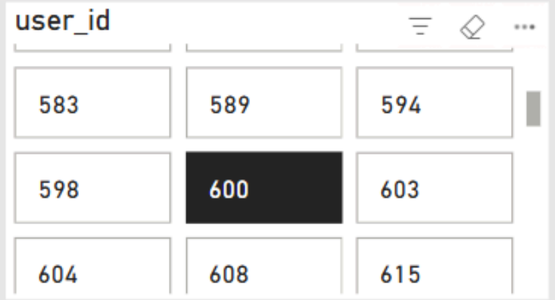
PHASE 3: CREATING A DASHBOARD

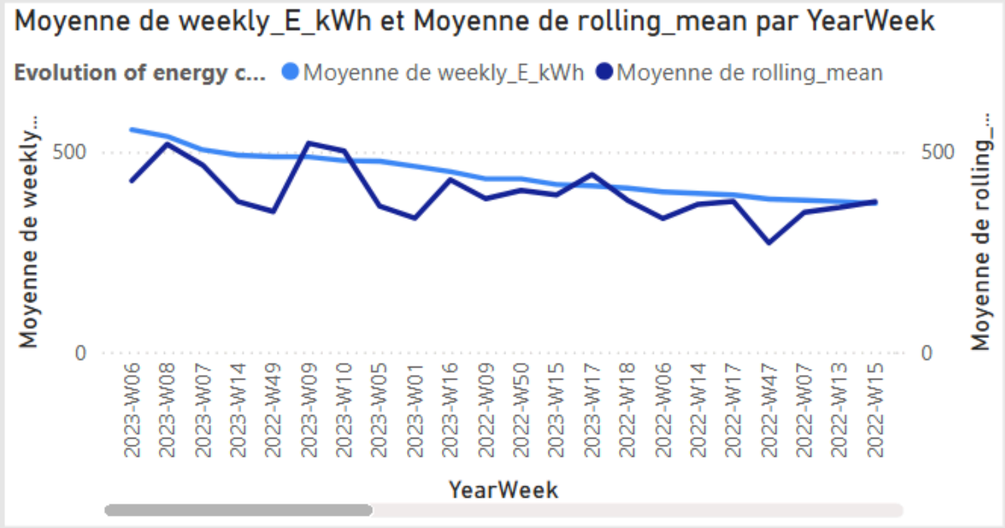
1. Estimated energy consumption

This graph shows how much electricity a user consumes each week, and compares it to a 4-week moving average.
This graph includes an interactive user filter (slicer), allowing us to explore the energy usage of each user individually over time.

- See trends (increasing or decreasing consumption)
- Detect seasonal patterns
- Spot sudden changes in usage

Energy Consumption Over Time





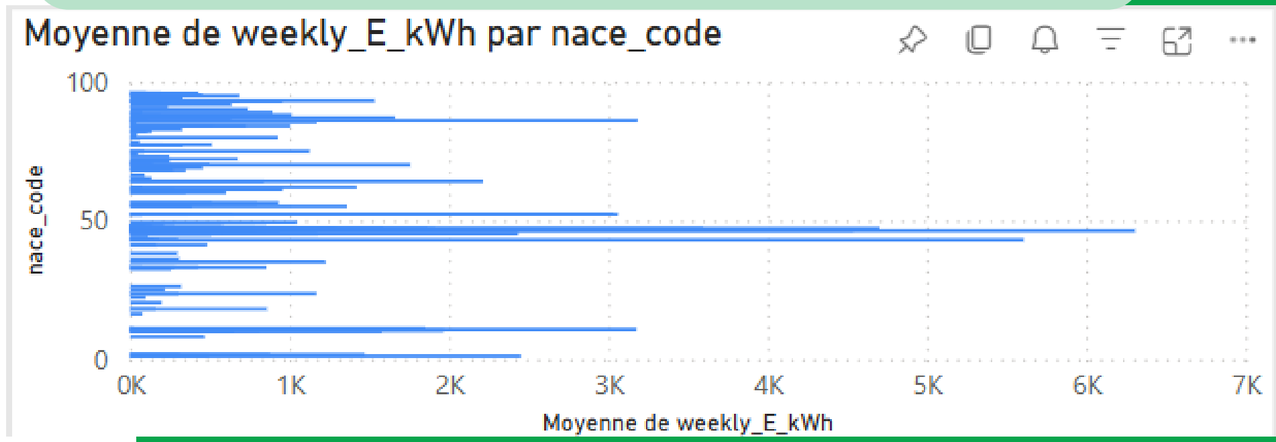
PHASE 3: CREATING A DASHBOARD

1. Estimated energy consumption

It shows the average weekly energy consumption per nace_code, which corresponds to the type of economic activity (industry, education, services, etc.).

- Compare energy usage by sector
- Identify which activity types consume the most electricity
- Help prioritize energy-saving strategies by industry

Energy Use by Activity Type



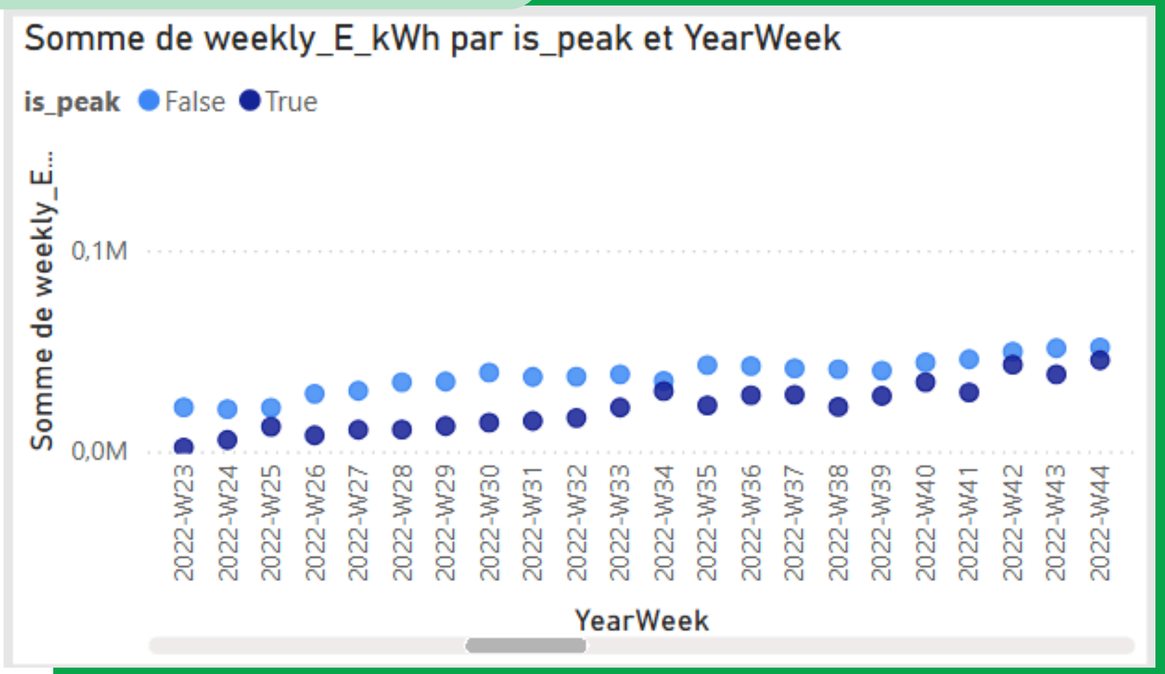
PHASE 3: CREATING A DASHBOARD

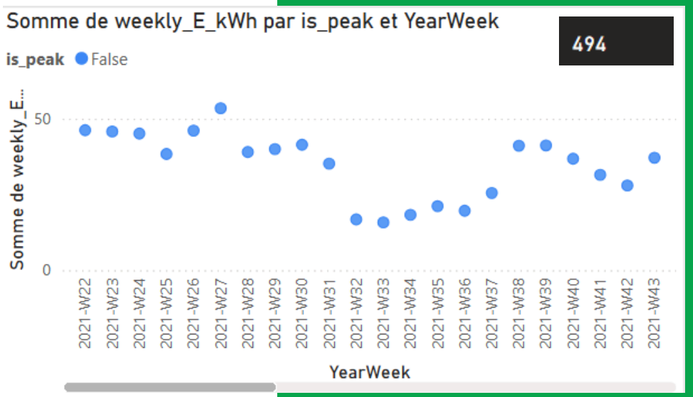
2. Detection of consumption peaks

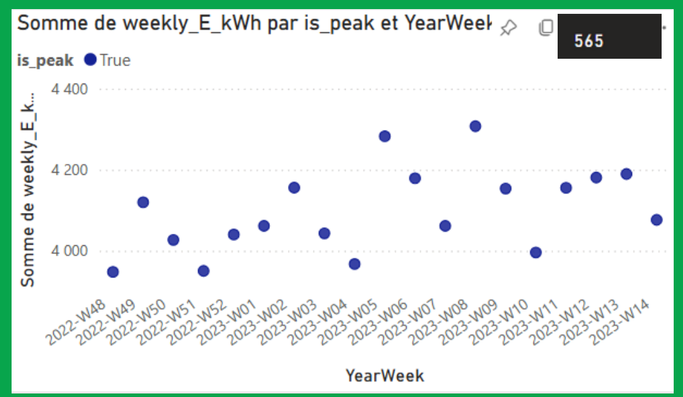
This graph displays only the weeks where energy peaks happened, as dots on a chart.

- Find when high consumption occurred
- Spot outliers and unusual behavior
- Understand which users or weeks are critical

Peak Detection







PHASE 3: CREATING A DASHBOARD

2. Detection of consumption peaks

This table lists all the weeks with a peak, along with key information:
user_id, site_id, YearWeek, nace_code, etc.

- Analyze peak cases in detail
- Identify sectors or buildings using too much energy
- Easily filter by user, year, or activity

Dynamic Table of Peaks

Dynamic peak table (peak is true)

user_id	YearWeek	weekly_E_kWh	site_id	nace_code	department_encoded
501	2023-W23	4 488,75	536	55,10	3
501	2023-W24	4 242,18	536	55,10	3
501	2023-W25	3 758,12	536	55,10	3
565	2022-W48	3 948,00	3843	47,30	5
565	2022-W49	4 119,87	3843	47,30	5
565	2022-W50	4 027,04	3843	47,30	5

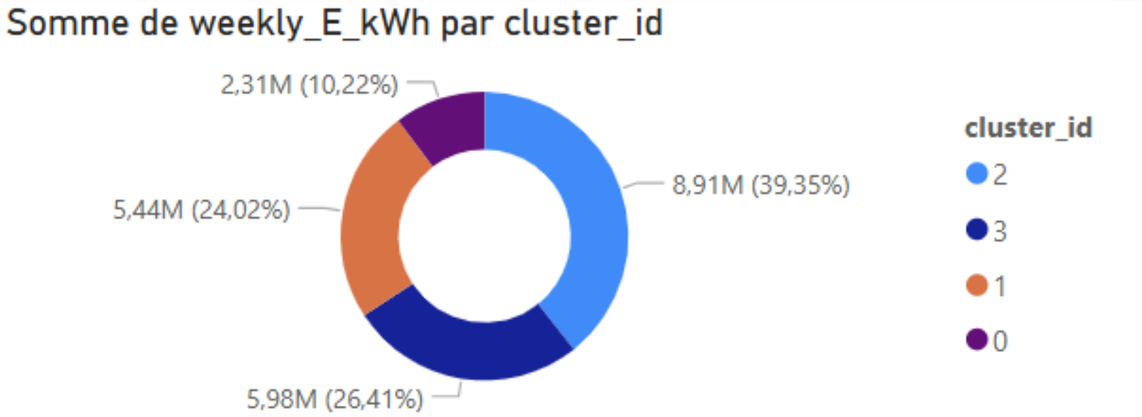
PHASE 3: CREATING A DASHBOARD

3. Classification or clustering results

This donut chart displays the total energy consumption per cluster, based on KMeans clustering.
Each cluster_id groups users with similar energy usage behavior, influenced by indoor conditions such as temperature and humidity.

- Understand how consumption is distributed across user groups
- See which cluster is the most energy-intensive
- Use it to target actions by behavioral profiles

Energy Consumption by Consumption Profile (Cluster)



PHASE 3: CREATING A DASHBOARD

In conclusion, this dashboard allowed us to explore and better understand energy consumption patterns through key visualizations.

We analyzed weekly consumption trends, detected consumption peaks, and identified user behavior clusters.

These insights can support targeted energy-saving strategies and better decision-making across sectors.

CONCLUSION: RECOMMANDATIONS

- **Target High-Consumption Activities :**
  Focusing on energy audits and efficiency improvements in high-demanding sectors
- **Regional Prioritization :**
  Targeted regional energy-saving programs, such as renewable energy incentives, or awareness campaigns in the most consuming areas(like Reunion Island)
- **Predictive Energy Monitoring :**
  integrate the Random Forest model into an energy monitoring system to forecast consumption in real time and identify unusual usage before it becomes costly.
- **Automated Anomaly Detection :**
  Using Isolation Forest and K-Means models to flag anomalies, such as sudden spikes or abnormal temperature readings
- **Encourage Energy Awareness :**
  Share model insights with operational teams to raise awareness

OUR TEAM

ISMAEL

LISA

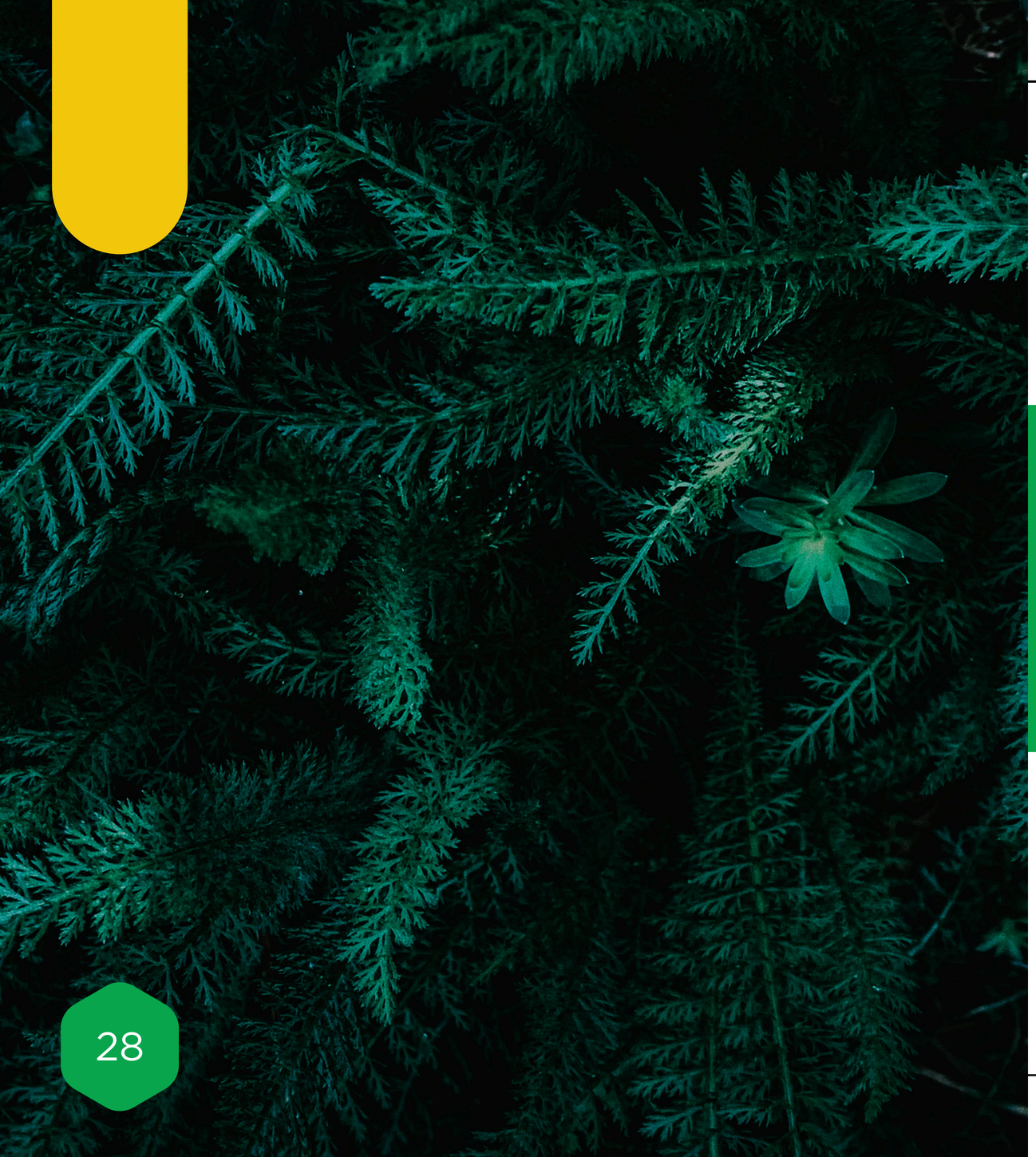
NGOULAYE

LEINA

KYLIE

LISA

DJIBRIL



THANK YOU
FOR YOUR ATTENTION !