Projet de Data Science

Analyse et Prevision de la Consommation Electrique

Réalisé par :

- Othmane Ouzzine
- Zaynab EL moudni
- Hajar Bouzid
- Hakkou mouataz

Encadré par : M. Yasser El Alami

Classe :1ème Année Master Ingénierie des Systèmes Informatiques (ISI) –Hybride

Établissement : SUPMTI RABAT

Année universitaire : 2025/2026

Le : 01/02/2026

Objectif du Projet

Ce projet de data science a pour but d'etudier et analyser les tendances de consommation energetique a partir de donnees reelles. Les objectifs principaux sont :

- Comprendre la structure des donnees de consommation electrique • Identifier les tendances et relations importantes entre les variables
- Construire des modeles predictifs pour prevoir la consommation totale (`totalkW_mean`) • Communiquer clairement les resultats et recommandations

Type de probleme : Regression (variable cible numerique continue)

Variable cible : `totalkW_mean` (consommation totale moyenne en kW)

Dataset

- **Source** : Donnees de consommation electrique quotidienne
- **Taille** : 600 lignes, 34 colonnes
- **Periode** : Juin 2018 a 2020
- **Variables** : Mesures de consommation de plusieurs compteurs (53, 71, 71A, 83), donnees meteorologiques (temperature, humidite, insolation), et variables temporelles

Structure du Projet

```
.
├── data/
│   ├── raw/                    # Donnees brutes
│   ├── processed/              # Donnees nettoyees et preparees
│   └── external/               # Donnees externes
├── notebooks/
│   ├── 01_collecte_donnees.ipynb
│   ├── 02_analyse_exploratoire_EDA.ipynb
│   ├── 03_preparation_donnees.ipynb
│   └── 04_modelisation.ipynb
├── src/                        # Modules Python reutilisables
│   ├── data_processing.py
│   ├── evaluation.py
│   ├── models.py
│   └── visualization.py
├── models/                     # Modeles entraines sauvegardes (.pkl)
├── visualizations/             # Graphiques generes
├── reports/                    # Rapports et resultats
├── presentation/               # Support de presentation
└── requirements.txt            # Dependances Python
```

1. Collecte des Donnees (`01_collecte_donnees.ipynb`)

Ce notebook est dedie au chargement initial et a l'exploration preliminaire du dataset. Il couvre :

- **Chargement du dataset** : Lecture du fichier CSV contenant 600 observations et 34 variables
- **Apercu des donnees** : Affichage des premieres lignes pour comprendre la structure
- **Informations generales** : Types de variables (22 float, 5 int, 7 string), utilisation memoire (0.34 MB)
- **Statistiques descriptives** : Resume statistique de toutes les variables numeriques
- **Analyse des valeurs manquantes** : Detection des colonnes avec des valeurs manquantes (`insolation` : 44.5%, `totalkW_w-1` et `totalkW_w/1` : 1.17%, `totalkW_d-1` et `totalkW_d/1` : 0.17%)
- **Verification des doublons** : Aucun doublon detecte
- **Cardinalite des variables categorielles** : Analyse du nombre de valeurs uniques par colonne
- **Sauvegarde** : Export du dataset brut pour les etapes suivantes

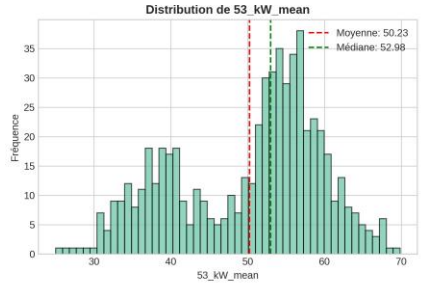
2. Analyse Exploratoire des Donnees - EDA (`02_analyse_exploratoire_EDA.ipynb`)

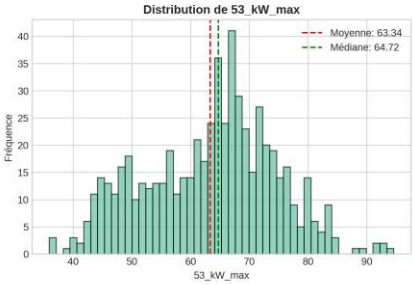
Ce notebook realise une analyse approfondie des donnees pour identifier des tendances et formuler des hypotheses. Il comprend :

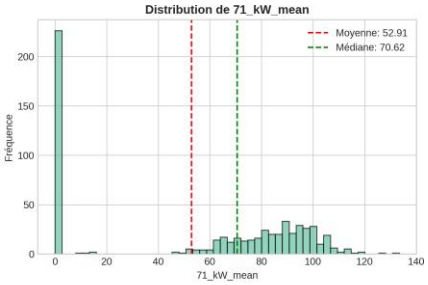
- **Statistiques descriptives avancees** : Calcul de la variance, asymetrie (skewness) et aplatissement (kurtosis) pour chaque variable
- **Analyse de l'asymetrie** : Identification des distributions asymetriques a droite (71A_kW_mean, HDD_25, NWD) et a gauche (RH)
- **Distributions des variables** : Histogrammes de toutes les variables numeriques avec moyenne et mediane
- **Detection des outliers** : Boxplots avec comptage des outliers par la methode IQR
- **Matrice de correlation** : Heatmap des correlations entre toutes les variables numeriques
- **Correlations fortes** : Identification de 48 paires de variables avec |r| > 0.7
- **Scatterplots** : Visualisation des paires les plus fortement correlees
- **Distribution des variables categorielles** : Analyse des variables temporelles (heures de pic)

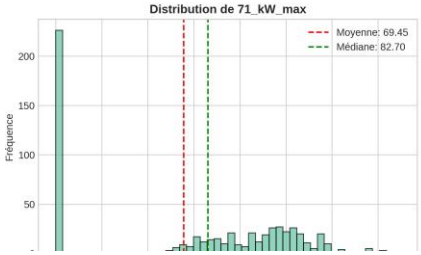
Visualisations

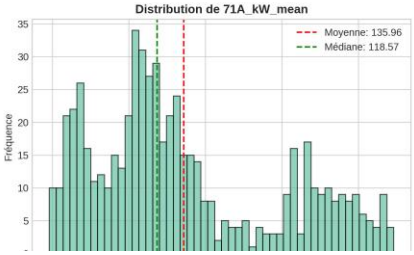
Distribution des variables numeriques :

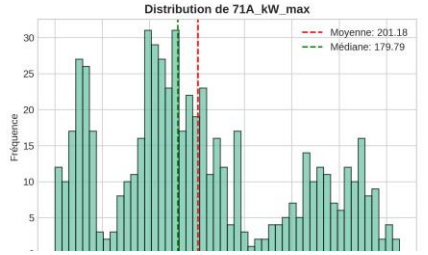

Distribution de 53_kW_mean
Moyenne: 50.23
Médiane: 52.98
Fréquence
53_kW_mean

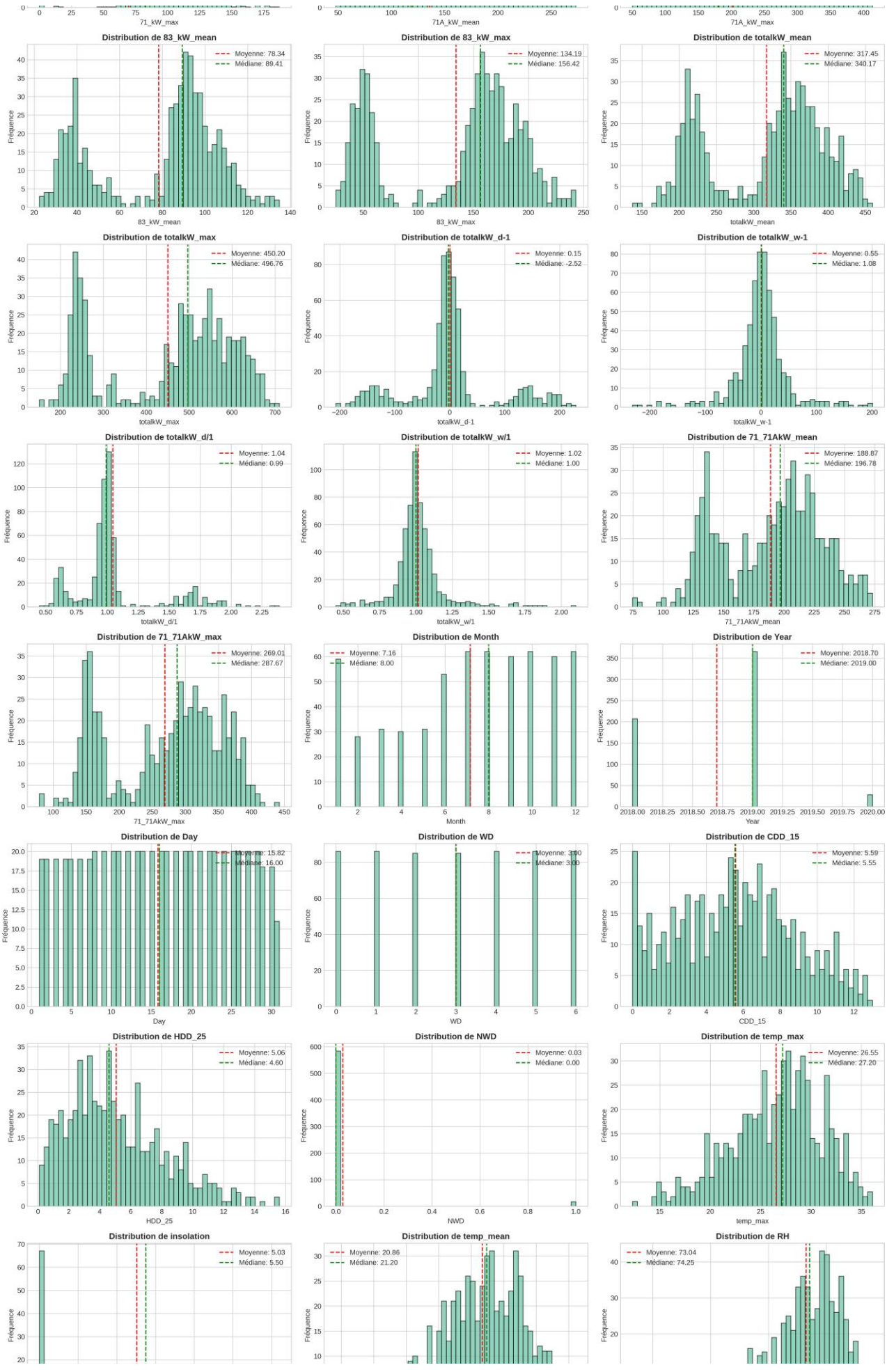

Distribution de 53_kW_max
Moyenne: 63.34
Médiane: 64.72
Fréquence
53_kW_max

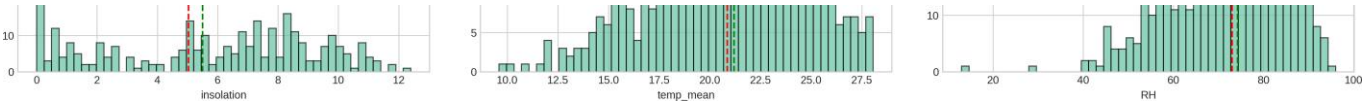

Distribution de 71_kW_mean
Moyenne: 52.91
Médiane: 70.62
Fréquence
71_kW_mean


Distribution de 71_kW_max
Moyenne: 69.45
Médiane: 82.70
Fréquence

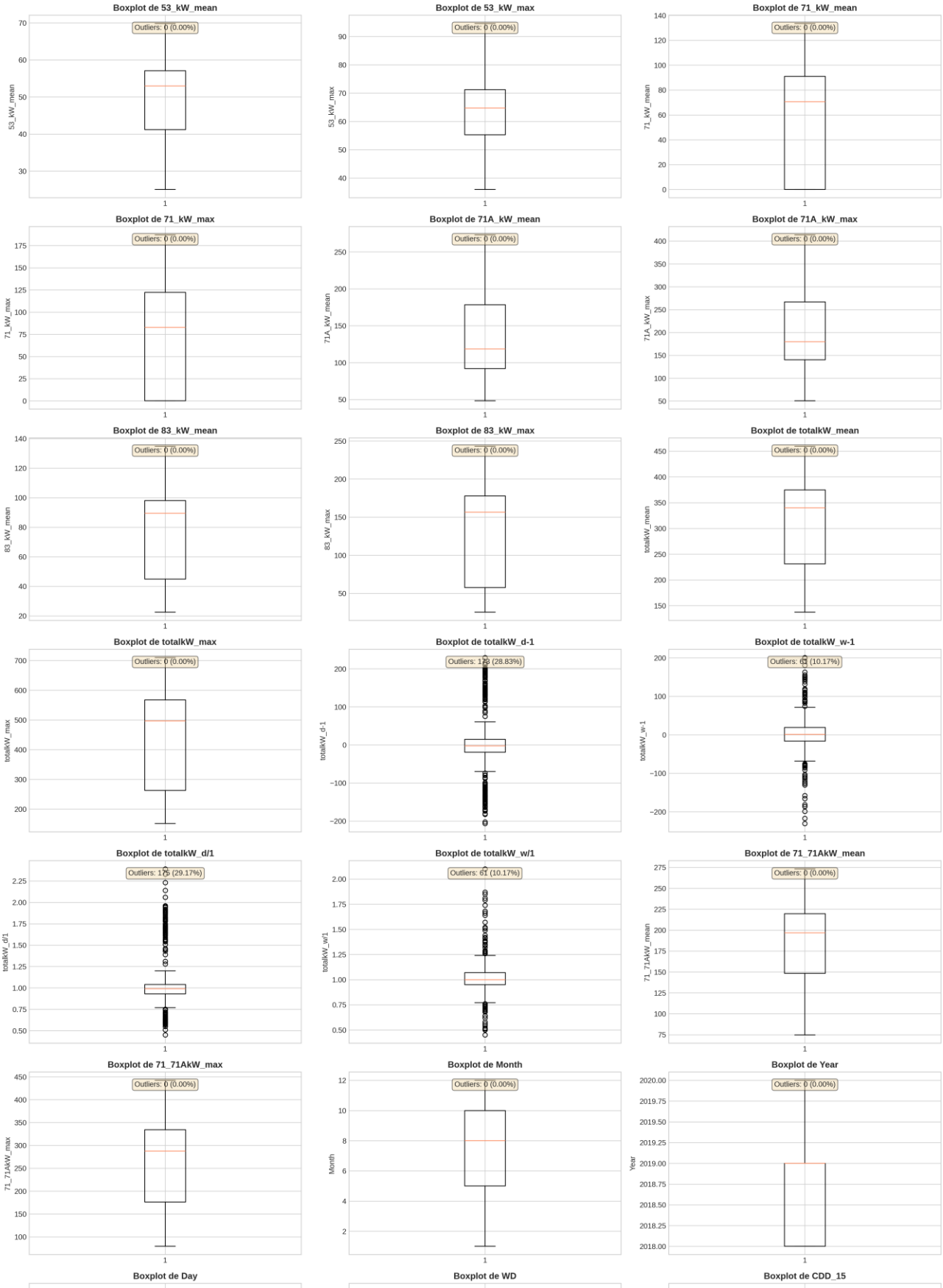

Distribution de 71A_kW_mean
Moyenne: 135.96
Médiane: 118.57
Fréquence

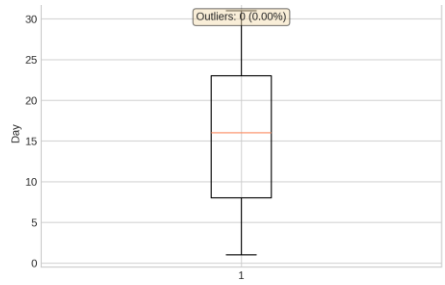

Distribution de 71A_kW_max
Moyenne: 201.18
Médiane: 179.79
Fréquence

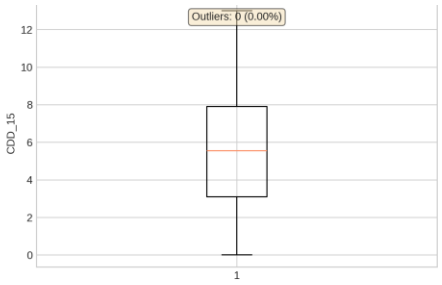

71_kW_max
71A_kW_mean
71A_kW_max
Distribution de 83_kW_mean
Moyenne: 78.34
Médiane: 89.41
83_kW_mean
Distribution de 83_kW_max
Moyenne: 134.19
Médiane: 156.42
83_kW_max
Distribution de totalkW_mean
Moyenne: 317.45
Médiane: 340.17
totalkW_mean
Distribution de totalkW_max
Moyenne: 450.20
Médiane: 496.76
totalkW_max
Distribution de totalkW_d-1
Moyenne: 0.15
Médiane: -2.52
totalkW_d-1
Distribution de totalkW_w-1
Moyenne: 0.55
Médiane: 1.08
totalkW_w-1
Distribution de totalkW_d/1
Moyenne: 1.04
Médiane: 0.99
totalkW_d/1
Distribution de totalkW_w/1
Moyenne: 1.02
Médiane: 1.00
totalkW_w/1
Distribution de 71_71AkW_mean
Moyenne: 188.87
Médiane: 196.78
71_71AkW_mean
Distribution de 71_71AkW_max
Moyenne: 269.01
Médiane: 287.67
71_71AkW_max
Distribution de Month
Moyenne: 7.16
Médiane: 8.00
Month
Distribution de Year
Moyenne: 2018.70
Médiane: 2019.00
Year
Distribution de Day
Moyenne: 15.82
Médiane: 16.00
Day
Distribution de WD
Moyenne: 3.00
Médiane: 3.00
WD
Distribution de CDD_15
Moyenne: 5.59
Médiane: 5.55
CDD_15
Distribution de HDD_25
Moyenne: 5.06
Médiane: 4.60
HDD_25
Distribution de NWD
Moyenne: 0.03
Médiane: 0.00
NWD
Distribution de temp_max
Moyenne: 26.55
Médiane: 27.20
temp_max
Distribution de insolation
Moyenne: 5.03
Médiane: 5.50
Distribution de temp_mean
Moyenne: 20.86
Médiane: 21.20
Distribution de RH
Moyenne: 73.04
Médiane: 74.25
Fréquence

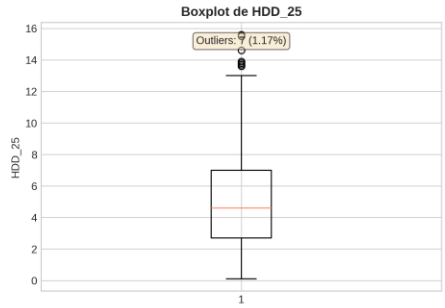


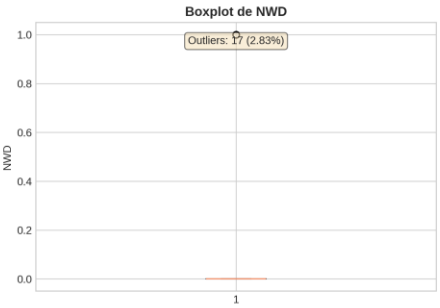
Boxplots pour la detection des outliers :

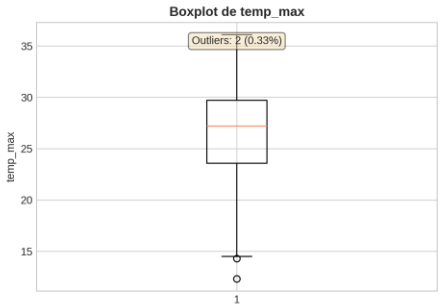


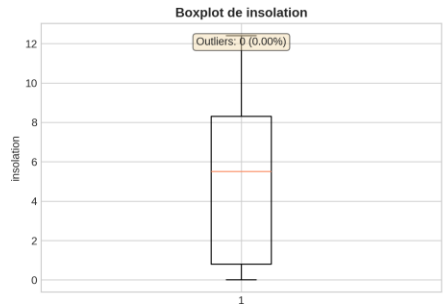


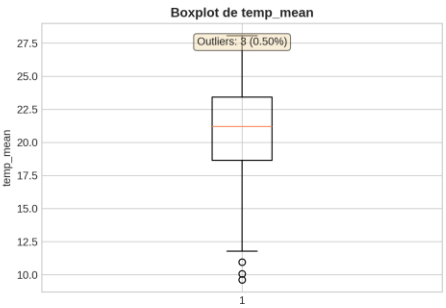


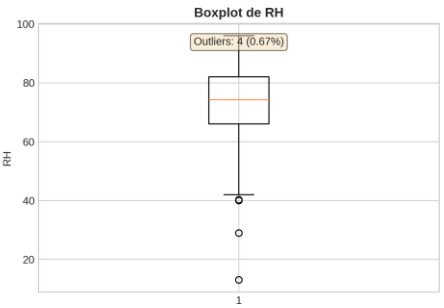




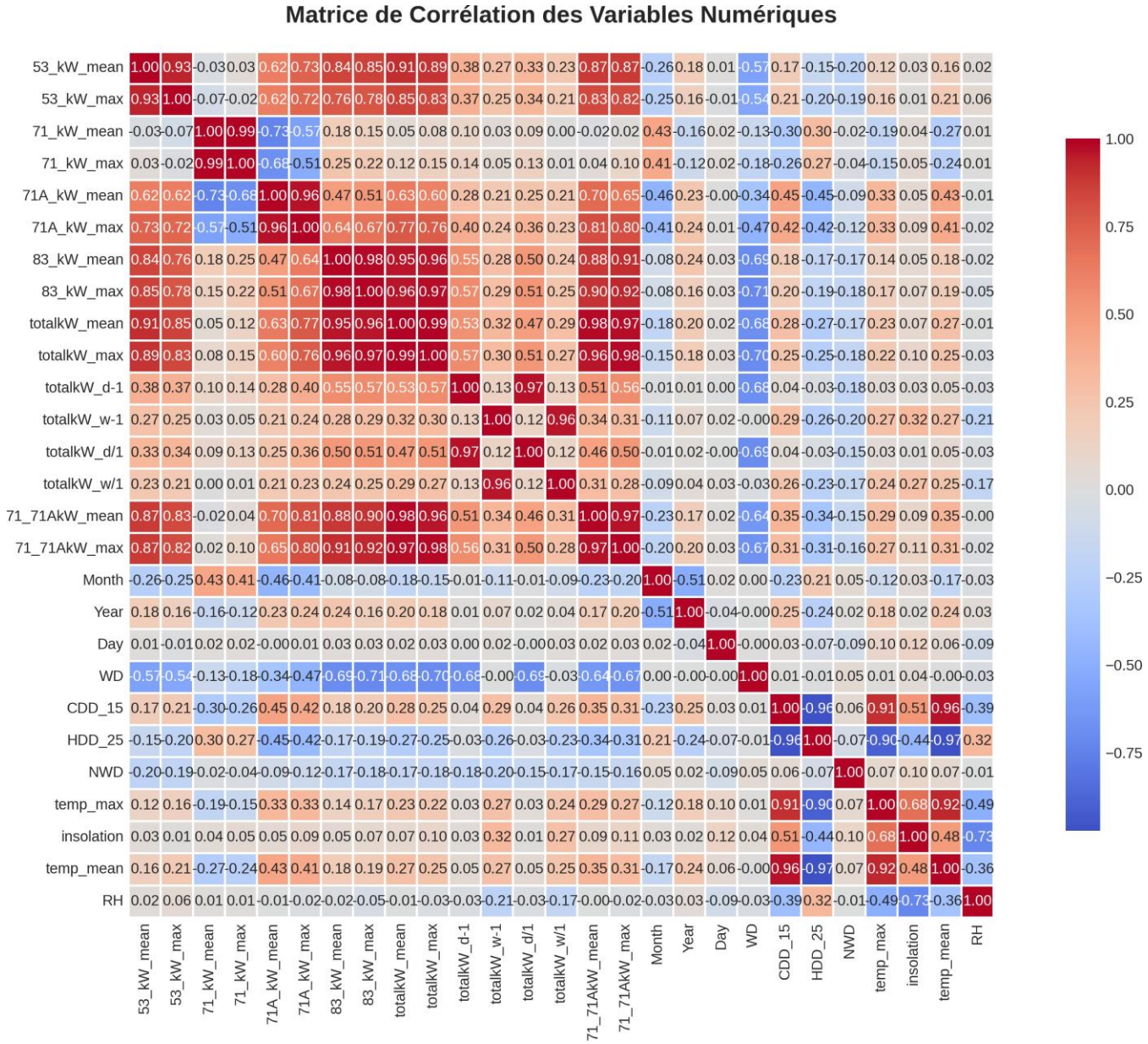




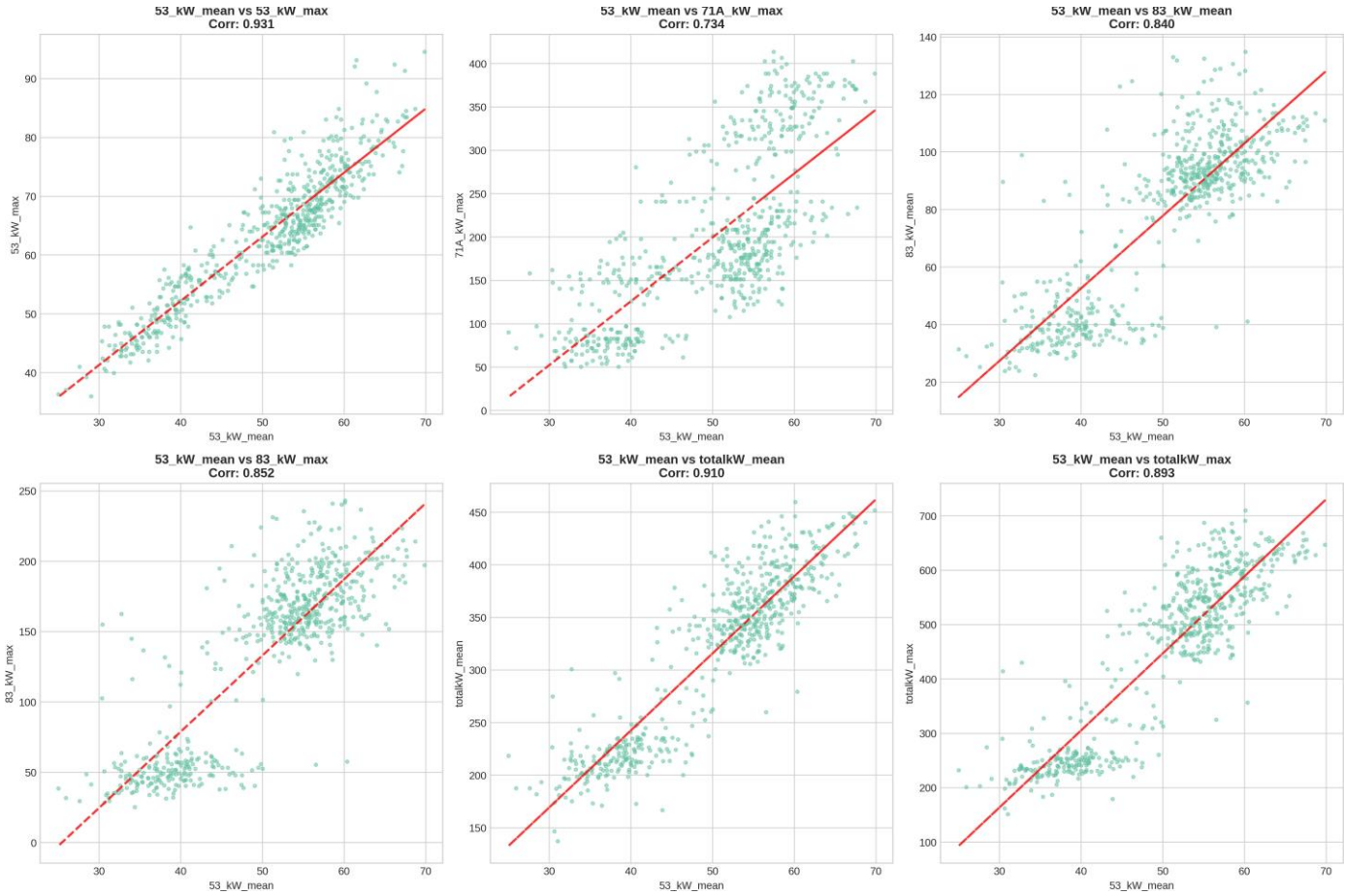




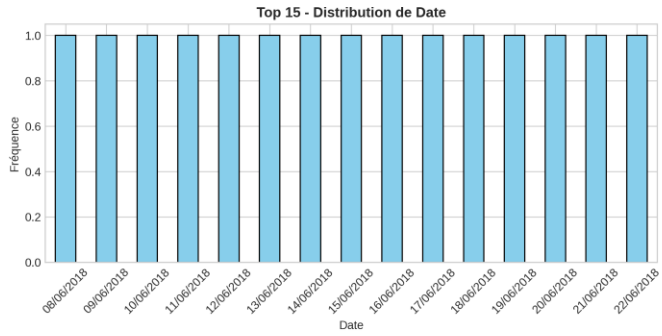
Matrice de correlation :

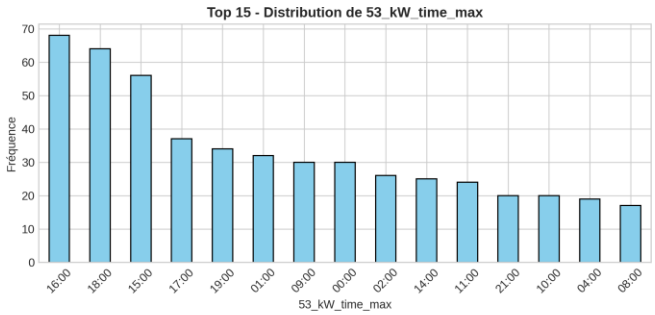


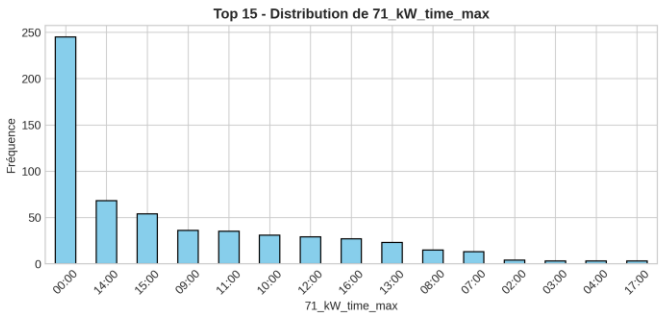
Scatterplots des correlations fortes :

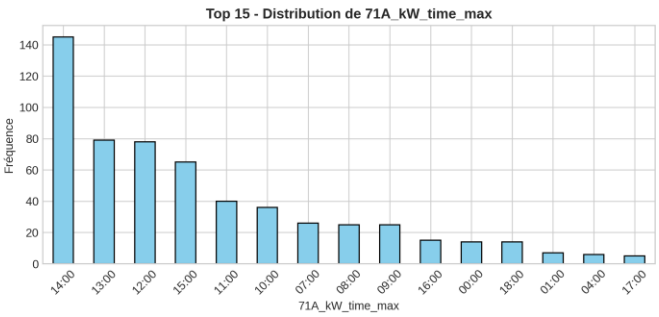


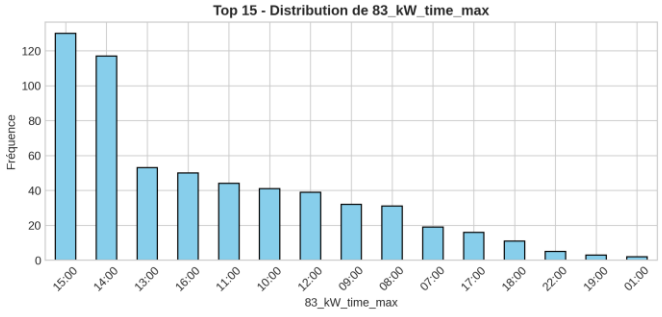
Distribution des variables categorielles :

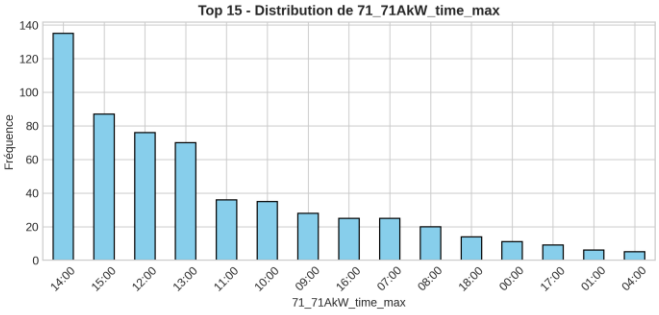


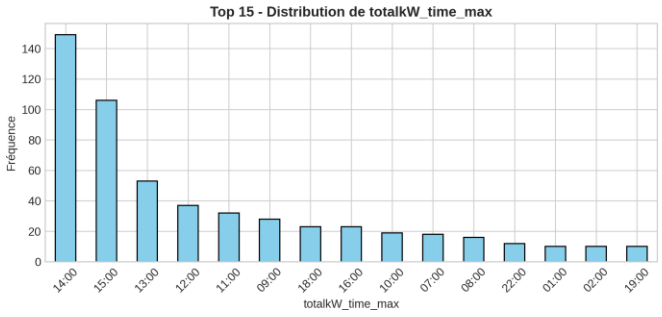












1. Preparation des Donnees (03_preparation_donnees.ipynb)

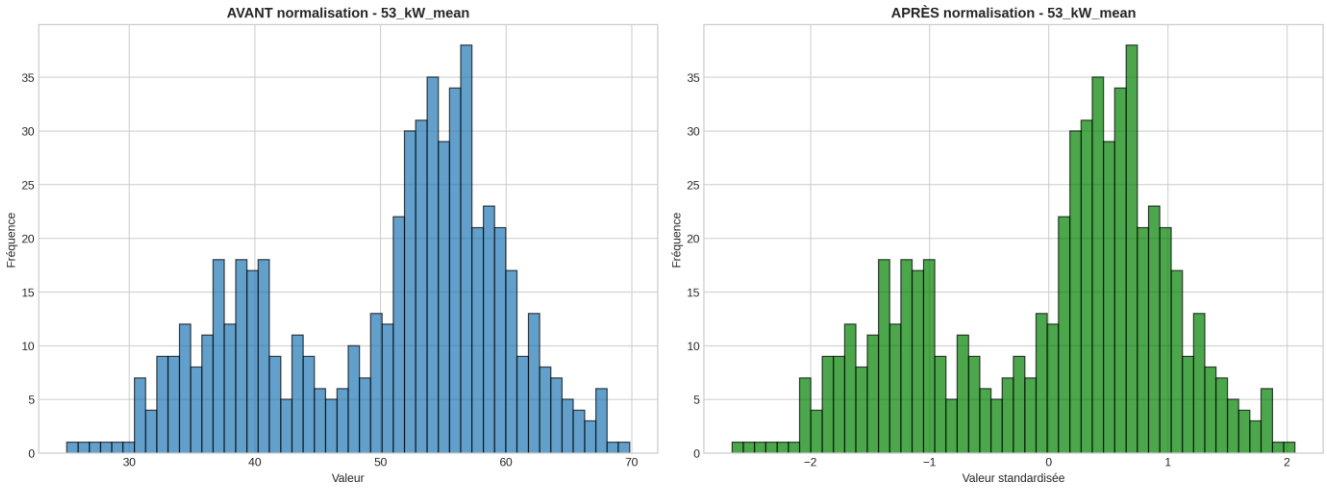
Ce notebook transforme les donnees brutes en un format exploitable pour la modelisation. Les etapes sont:

- **Suppression des doublons** : Verification et suppression (0 doublons trouves)
- **Gestion des valeurs manquantes** :Imputation par la mediane pour les variables numeriques (totalkW_d-1, totalkW_w-1, totalkW_d/1, totalkW_w/1, insolation)
- **Traitement des outliers :** Winsorization (plafonnement) de 503 valeurs extremes sur 9 variables avec la methode IQR
- **Encodage des variables categorielles** : Suppression des colonnes temporelles (Date, heures de pic) car trop de categories uniques
- **Normalisation :** Standardisation (Z-score) de toutes les features avec StandardScaler
- **Separation train/test** : 80% entrainement (480 lignes) / 20% test (120 lignes)

- **Sauvegarde** : Export des ensembles X_train, X_test, y_train, y_test et du scaler

Visualisation

Comparaison avant/apres normalisation :



3. Modelisation (`04_modelisation.ipynb`)

Ce notebook entraine et evalue trois modeles de regression pour predire la consommation electrique totale. Il comprend :

- **Modele 1 - Regression Lineaire** : Modele de base avec validation croisee 5-fold
- **Modele 2 - Random Forest Regressor** : Modele d'ensemble avec 100 arbres, profondeur max 15
- **Modele 3 - Gradient Boosting Regressor** : Modele de boosting avec taux d'apprentissage 0.1
- **Evaluation** : Metriques R², RMSE, MAE et MAPE pour chaque modele
- **Comparaison** : Tableau et graphiques comparatifs des performances
- **Importance des variables** : Identification des features les plus influentes pour chaque modele
- **Analyse des residus** : Distribution des erreurs de prediction
- **Sauvegarde** : Export des modeles entraines au format `.pkl`

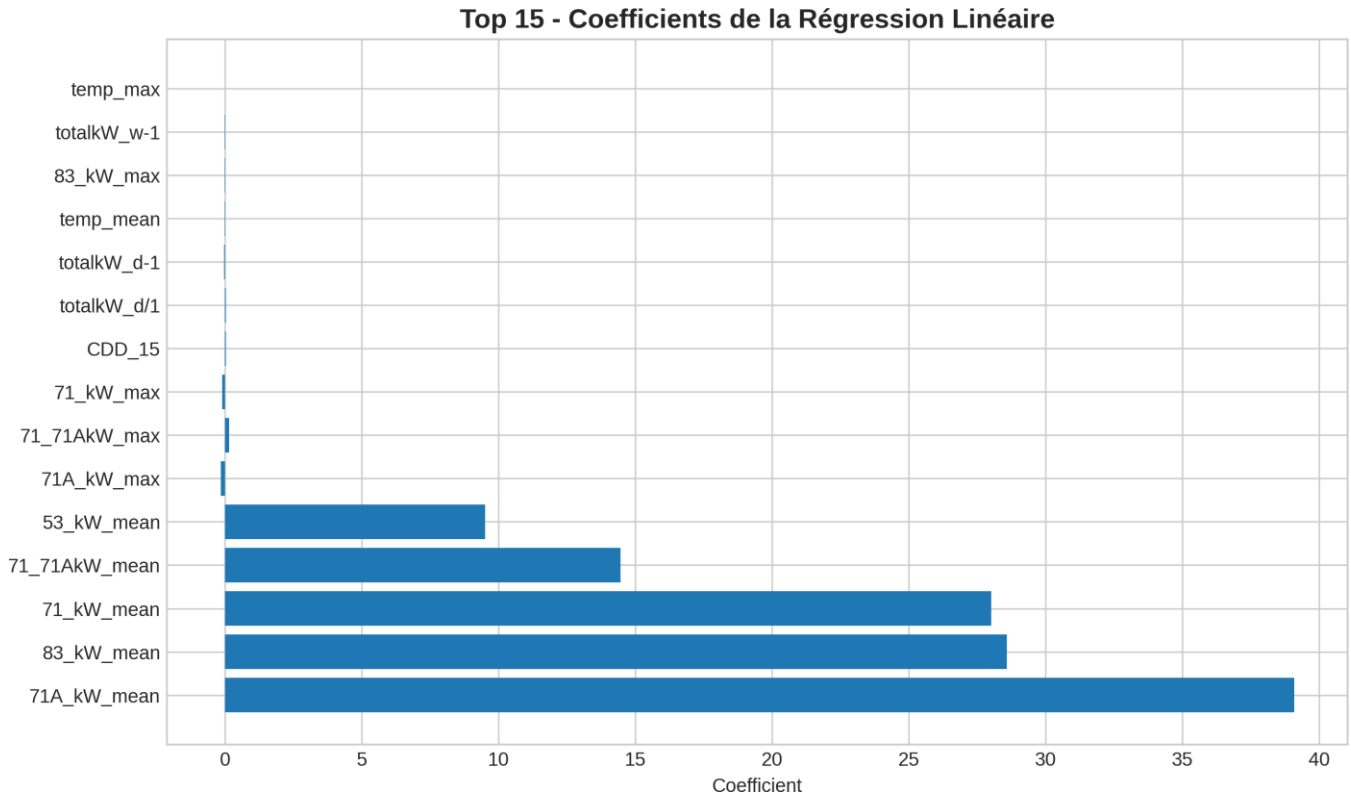
Resultats

Modele	R²	RMSE	MAE	MAPE (%)
Linear Regression	1.0000	0.0828	0.0662	0.02
Random Forest	0.9934	6.0606	4.1775	1.38
Gradient Boosting	0.9960	4.7214	3.2165	1.07

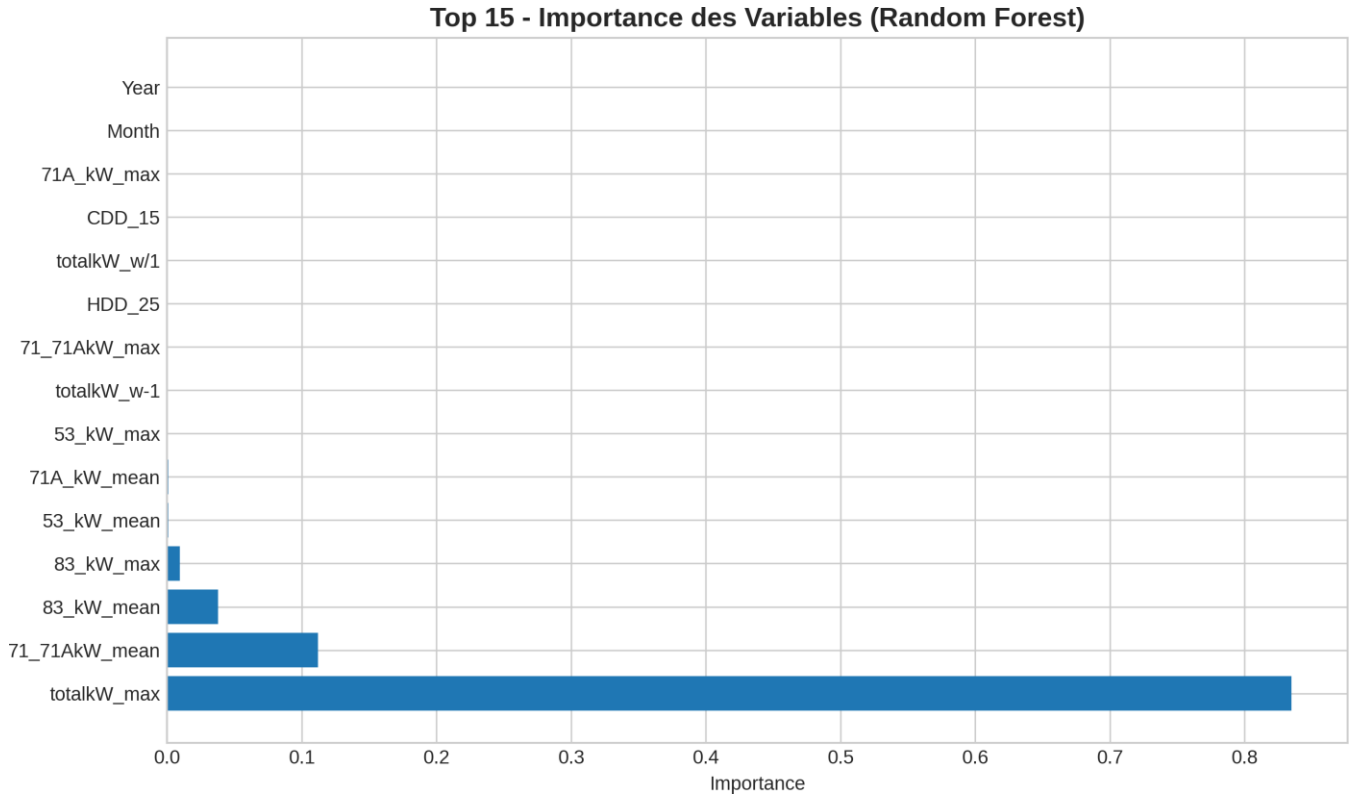
Meilleur modele : Regression Lineaire (R² = 1.0000)

Visualisations

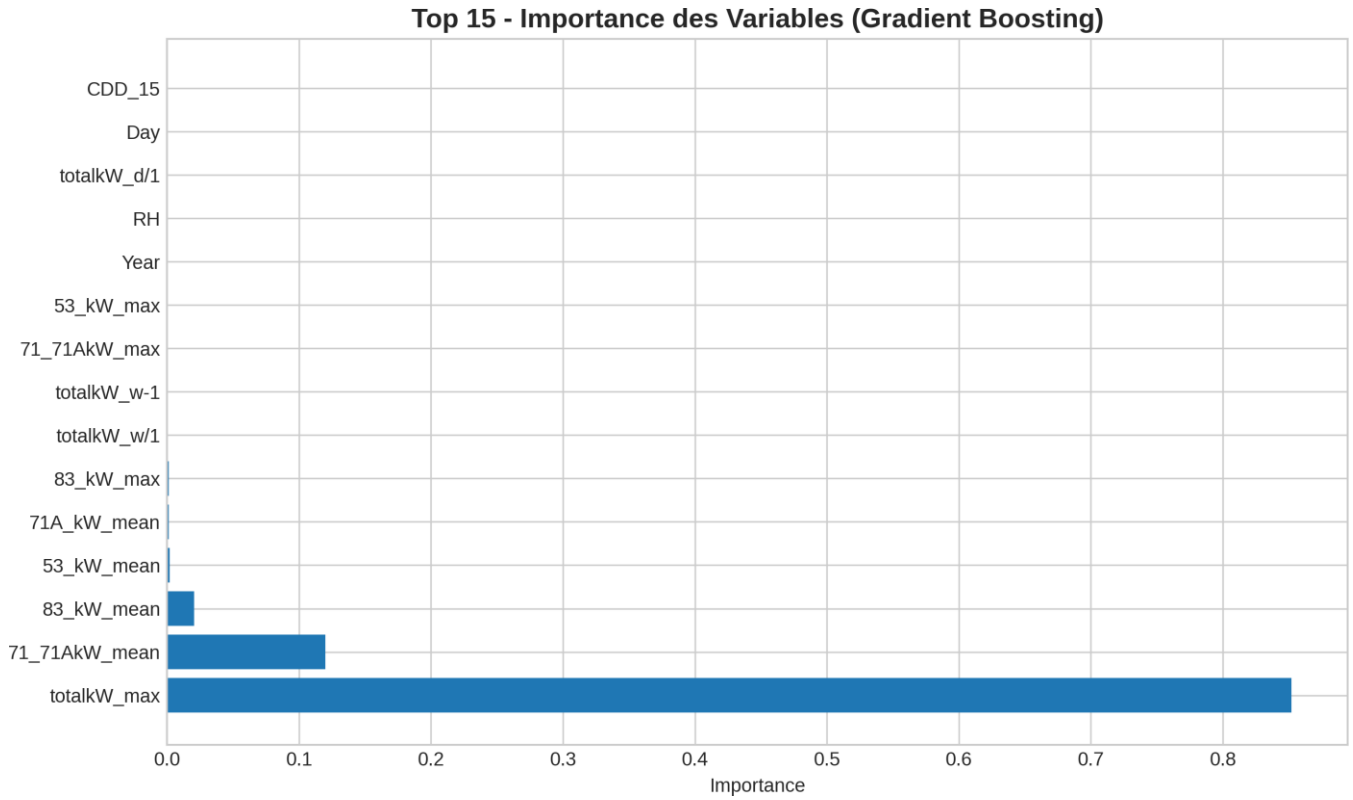
Coefficients de la regression lineaire :



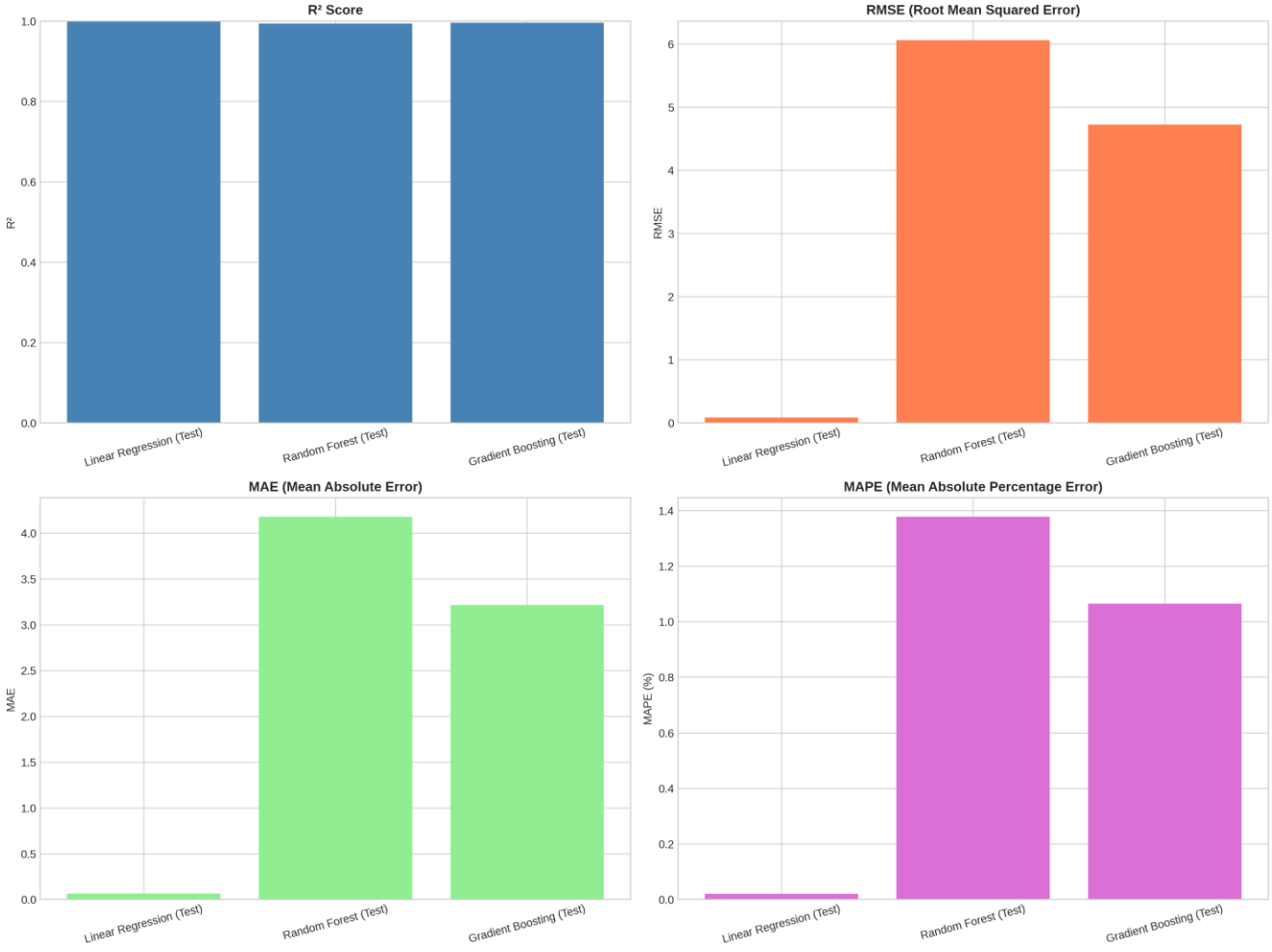
Importance des variables - Random Forest :



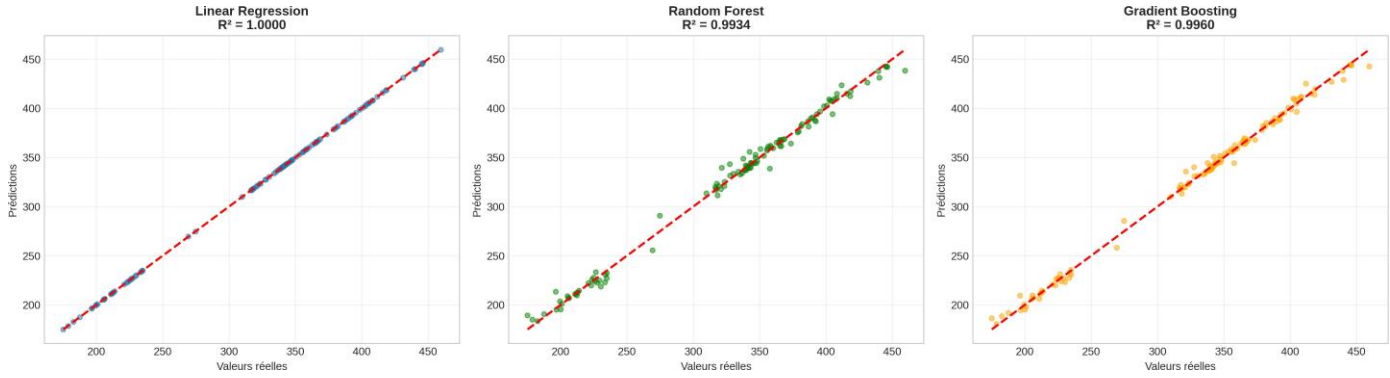
Importance des variables - Gradient Boosting :



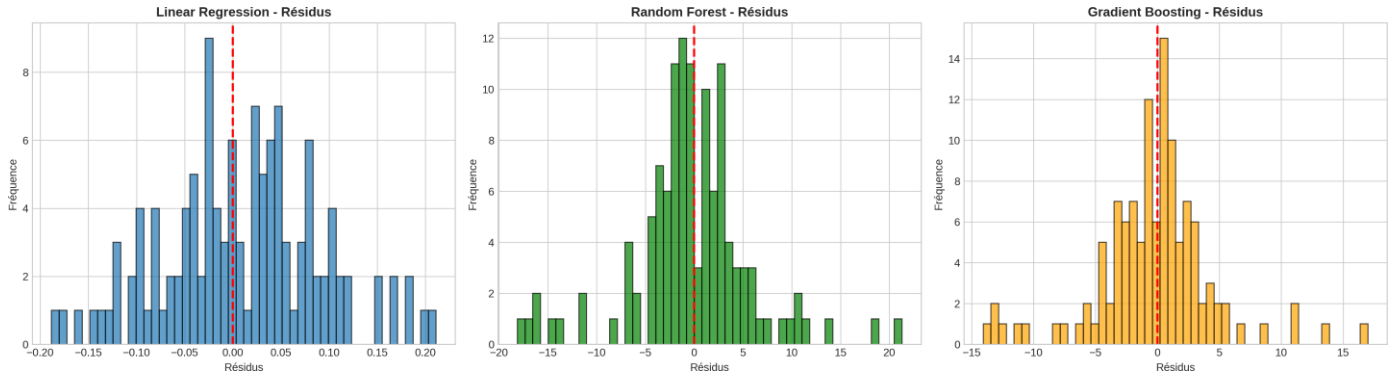
Comparaison des modeles :



Predictions vs Valeurs reelles :



Distribution des residus :



Technologies Utilisees

- **Python 3**
- **Pandas** / **NumPy** : Manipulation et traitement des donnees
- **Matplotlib** / **Seaborn** : Visualisations graphiques
- **Scikit-learn** : Modelisation (LinearRegression, RandomForestRegressor, GradientBoostingRegressor)
- **SciPy** : Analyses statistiques
- **Joblib** : Sauvegarde des modeles

Installation

```
pip install -r requirements.txt
```

Utilisation

Executer les notebooks dans l'ordre :

1. `01_collecte_donnees.ipynb` - Chargement et exploration initiale
2. `02_analyse_exploratoire_EDA.ipynb` - Analyse exploratoire et visualisations
3. `03_preparation_donnees.ipynb` - Nettoyage et preparation des donnees
4. `04_modelisation.ipynb` - Entrainement et evaluation des modeles