Project Presentation
https://github.com/shuawoo/EL_Gr3_T1

ELECTRICITY PRICE EXPLANATION

Jiuduo WANG
Shuya WU
Yingxuan WU
Yiru YIRU

OVERVIEW

01 INTRODUCTION & MOTIVATION

02 DATA INSPECTION

03 PREPROCESSING

04 FEATURE ENGINEERING

05 MODEL

06 HYPERPARAMETER TUNING

07 EVALUATION

Introduction & Motivation

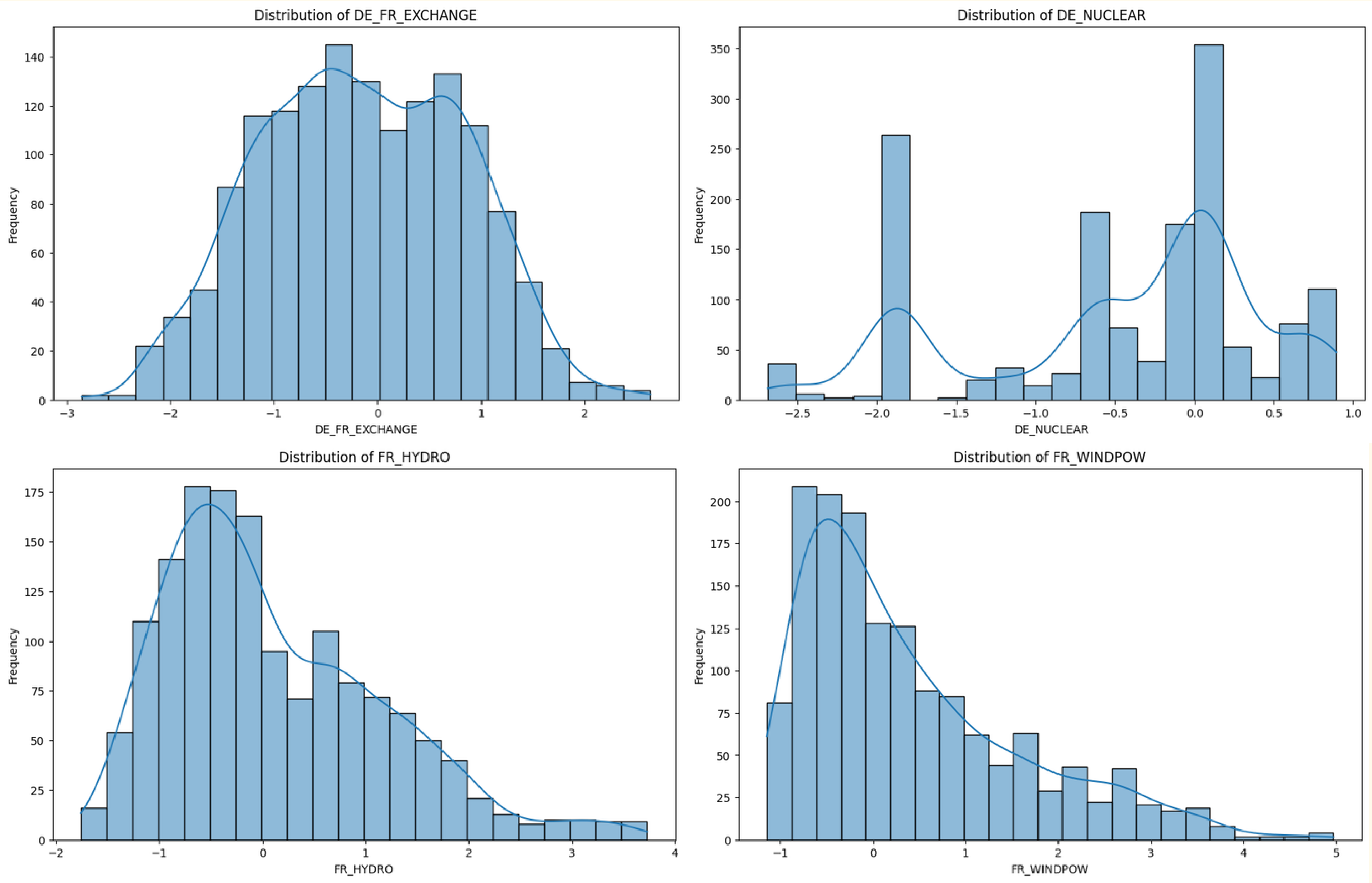
This project is about building models to explain the electricity price focusing on 2 specific countries: France and Germany. The modelisation in these country will be our goal.

Electricity price forecasting plays a crucial role in various aspects including energy trading, policy-making, household issue, and infrastructure planning.

Electricity price is influenced by multiple factors including local weather, global warming, wars, and government policy, which makes the electricity price model very complex.

The ensemble learning methods employed combine and aggregate advatanges of different algorithms, helping build a more robust explanation model by efficiently recognizing the electricity price dynamics.

DATA INSPECTION



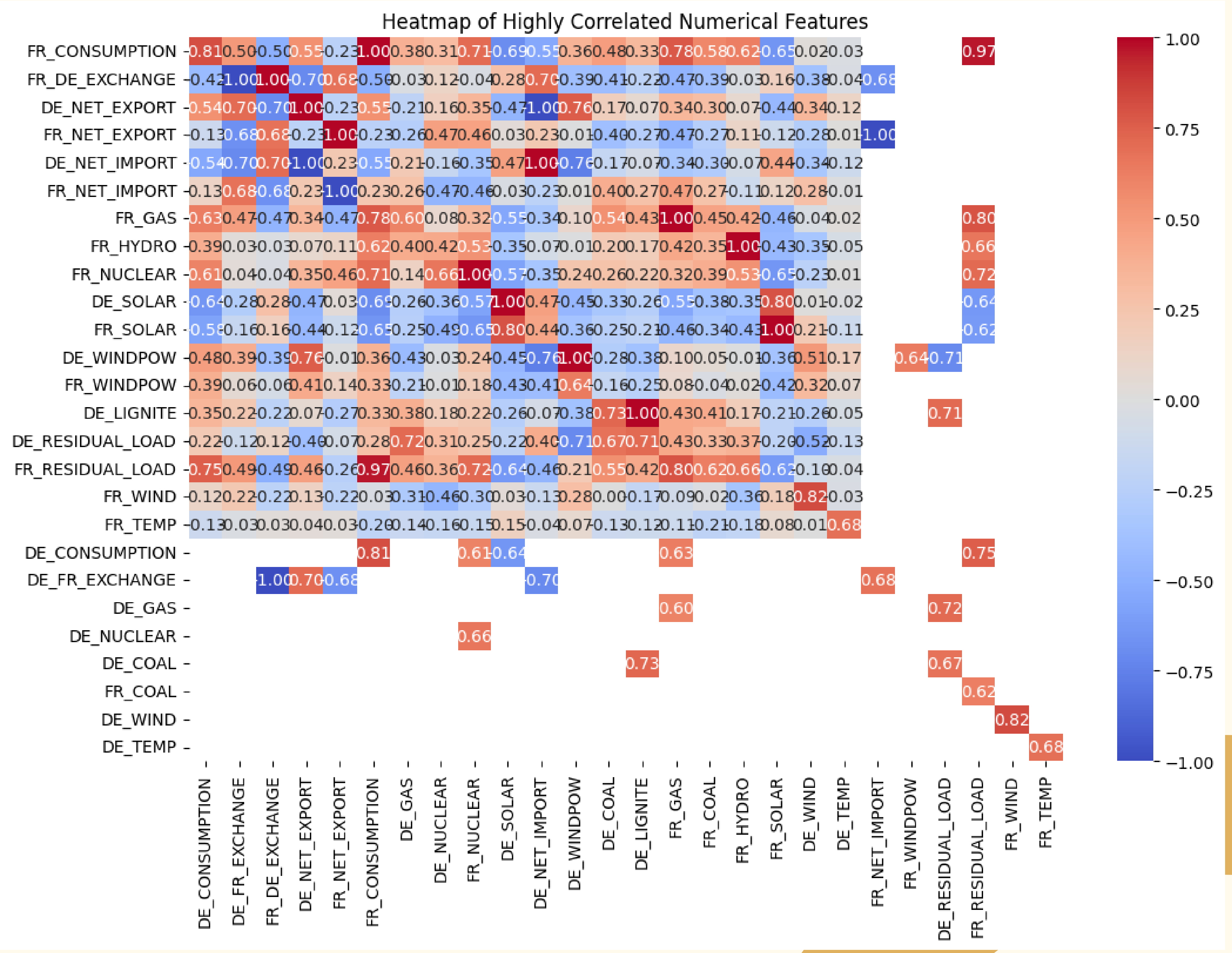
- The histograms are bell-shaped and closely resemble the **normal distribution**, which is an indication that the data may not need transformations for preprocessing.

- Some histograms show **potential outliers**, which will be considered in the data preprocessing stage.

DATA INSPECTION

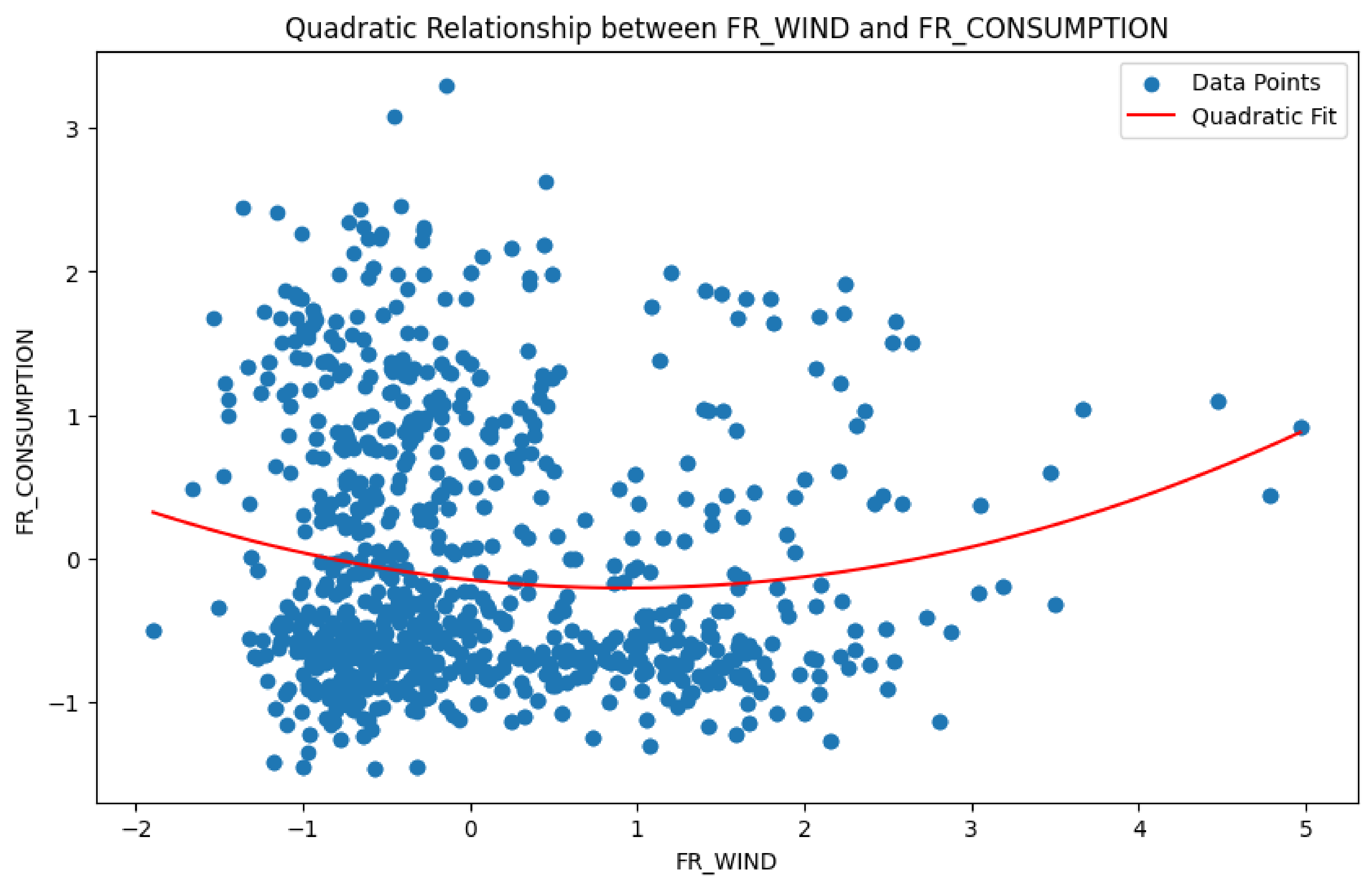
- There are **strong positive correlations** within each country's features

- The dataset exhibits **Cross-country correlations**, indicating possible interdependencies or similar patterns in energy usage or production between the two countries (which makes realistic sense since two countries are geographically close)

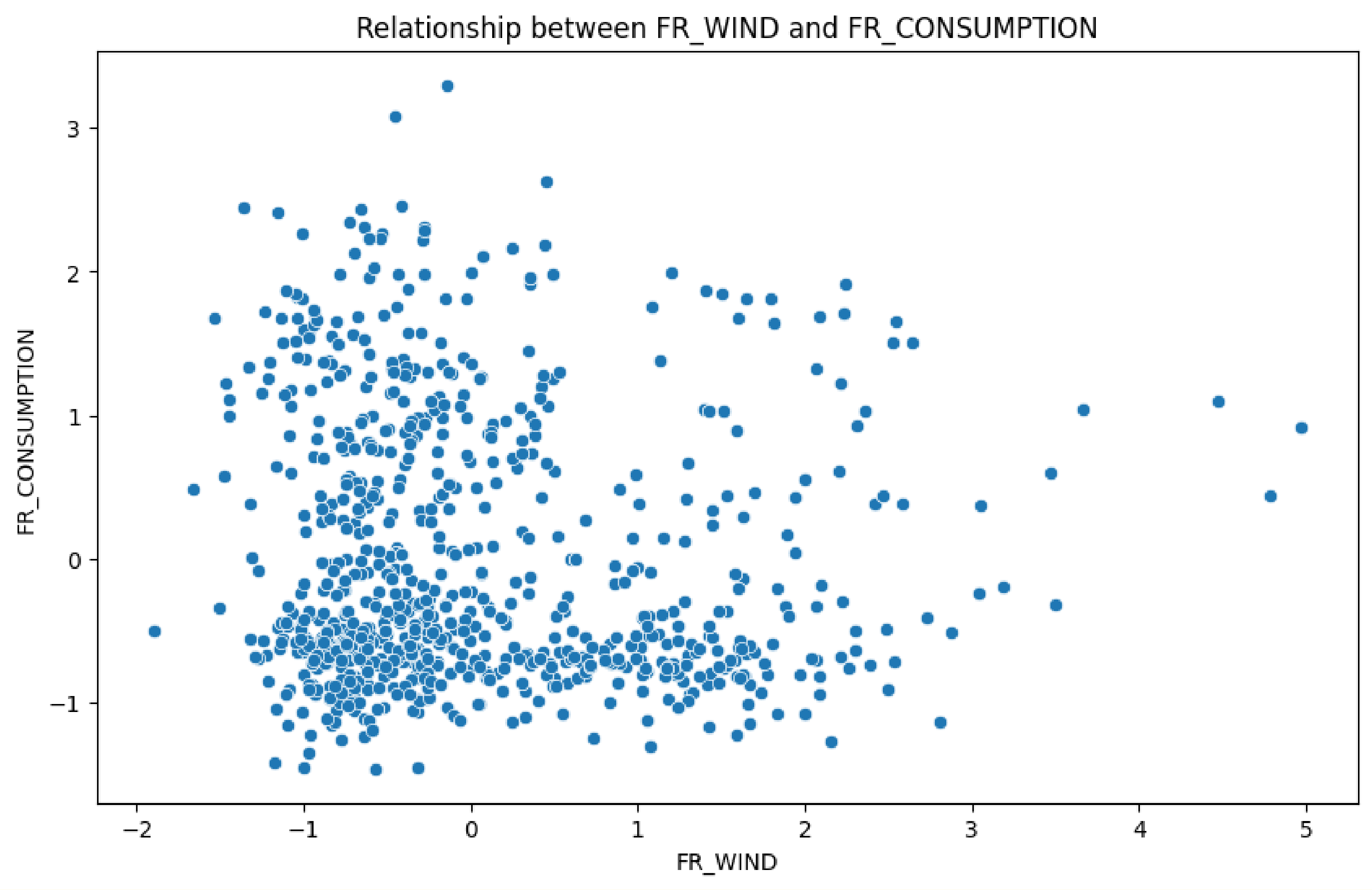


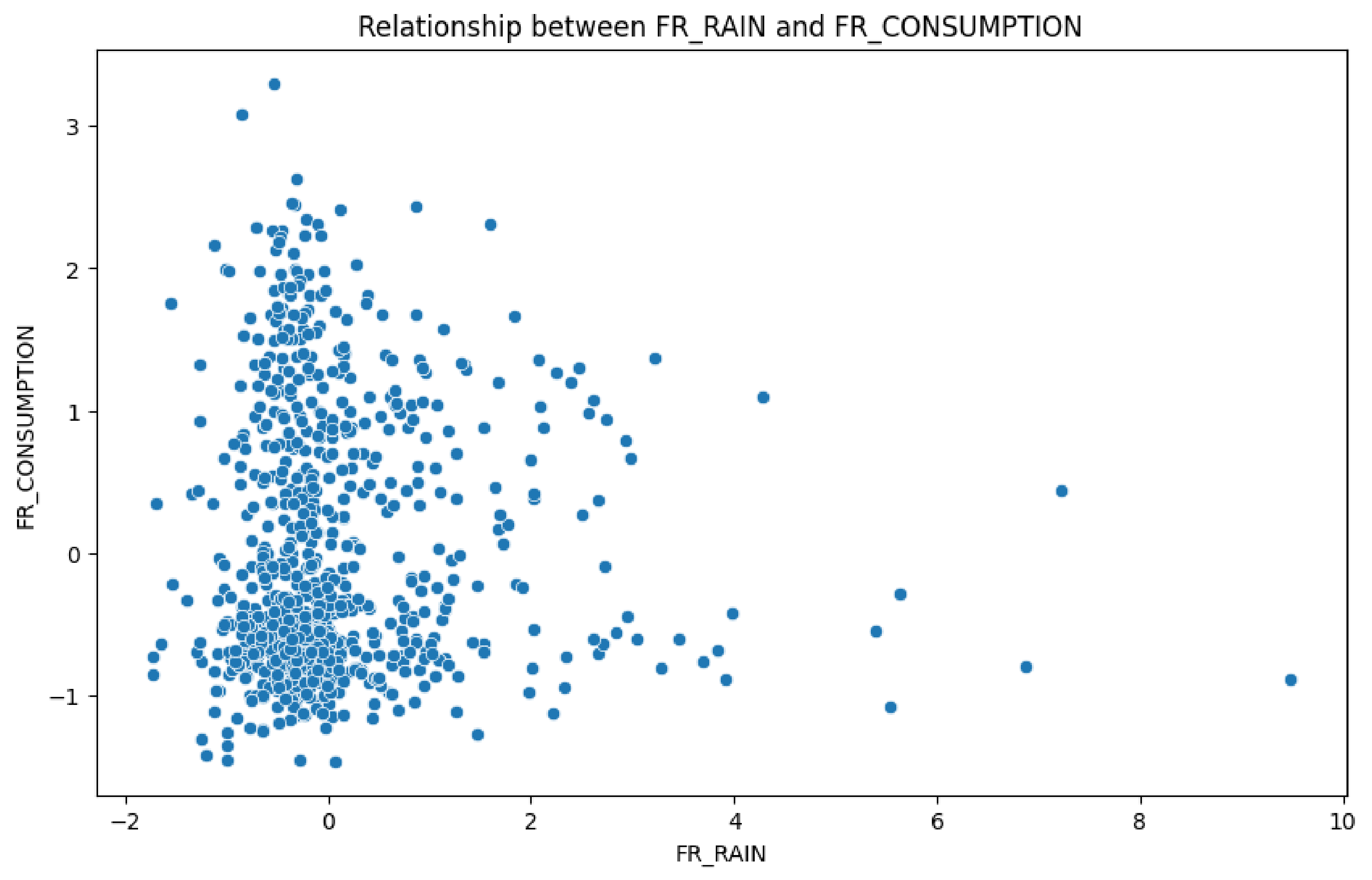
*: only highly correlated pairs are shown with threshold = 0.6

DATA INSPECTION



- There is **no strong quadratic relationships** between the variables

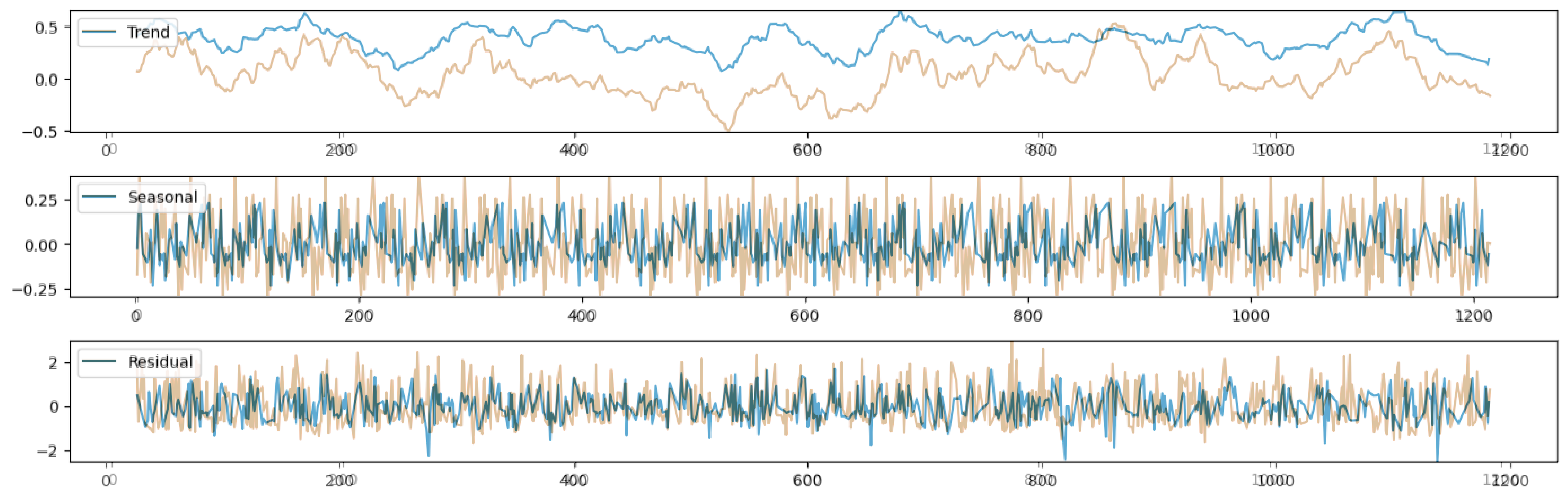




DATA INSPECTION

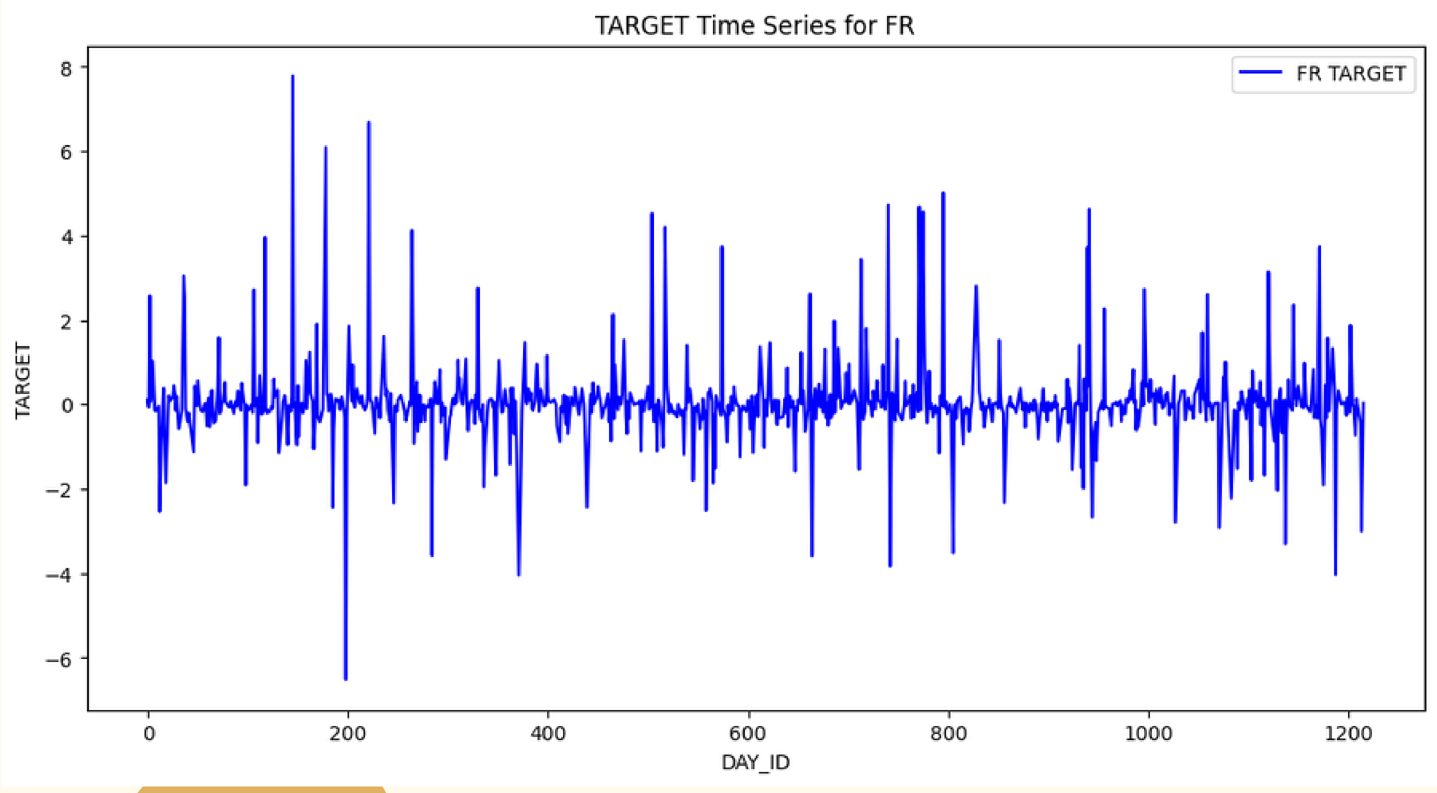
Divide the dataset into FR and DE

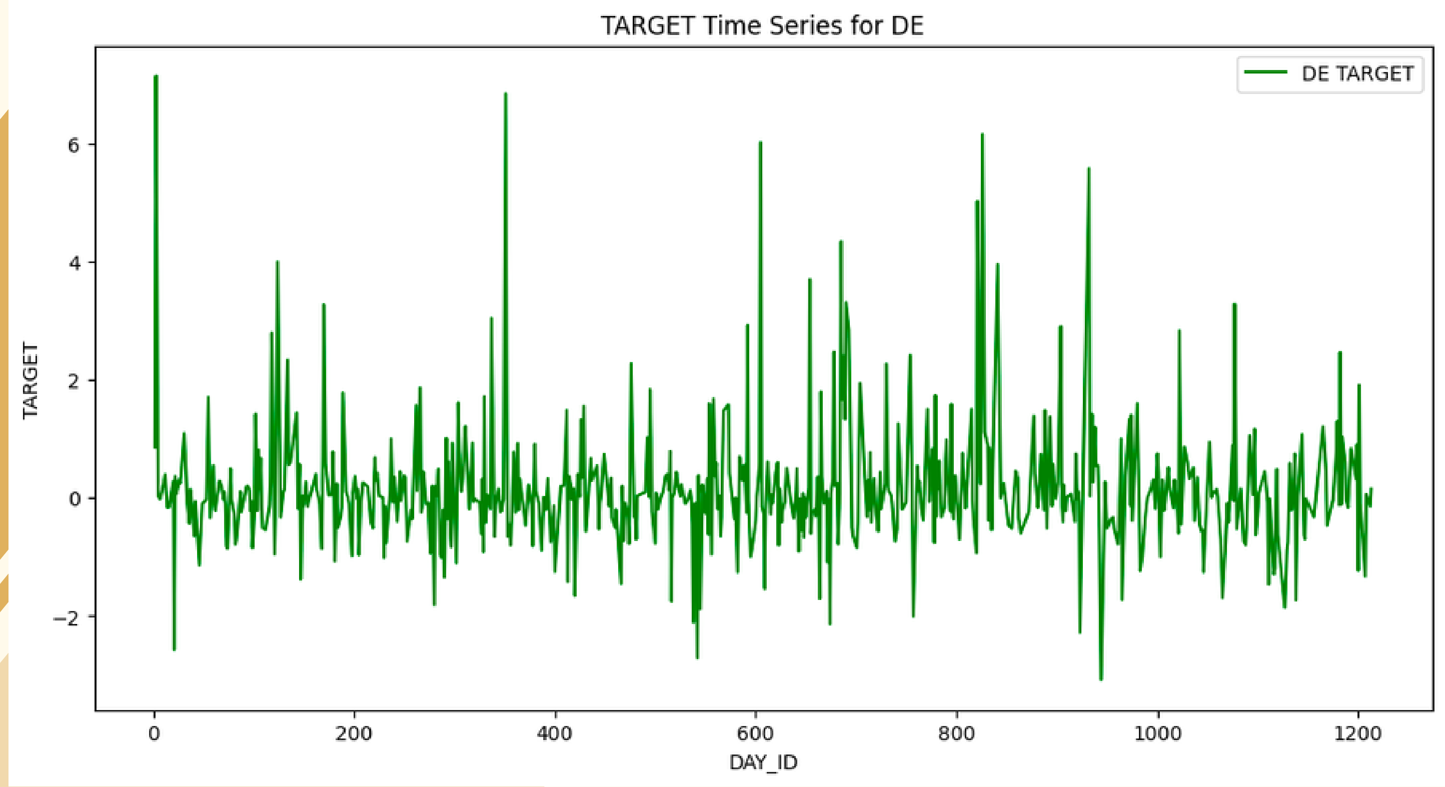
We first divide the dataset into French and German parts, and these two parts will be trained separately in the following process, because we observe different patterns between these two countries.



*: graph shown above is the overlaying of `FR_CONSUMPTION` and `DE_CONSUMPTION` using 30 as period, we can find that both variables have distinct trends, hence we generalize this idea and decide to divide the dataset into two section

PREPROCESSING





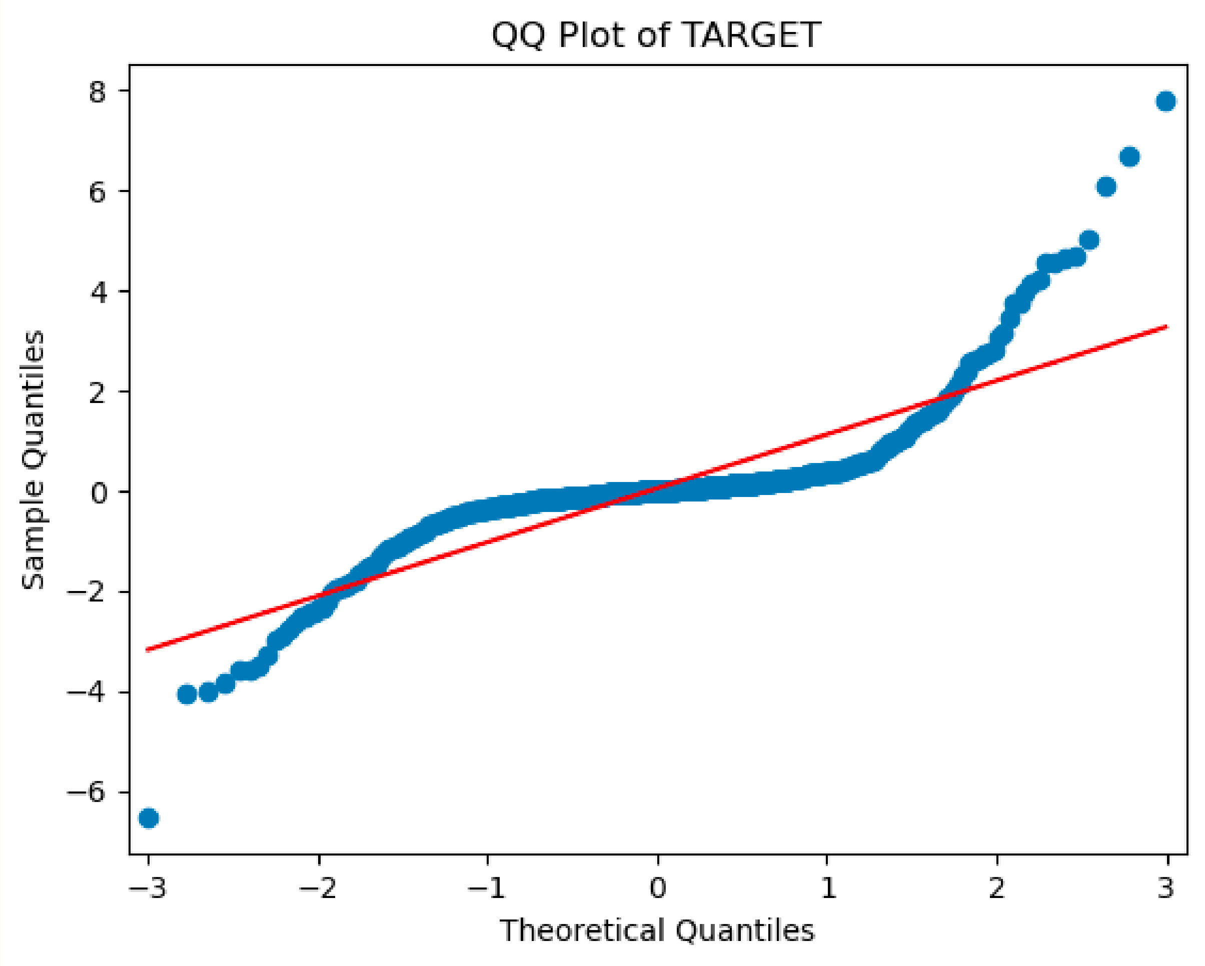
NULL VALUE

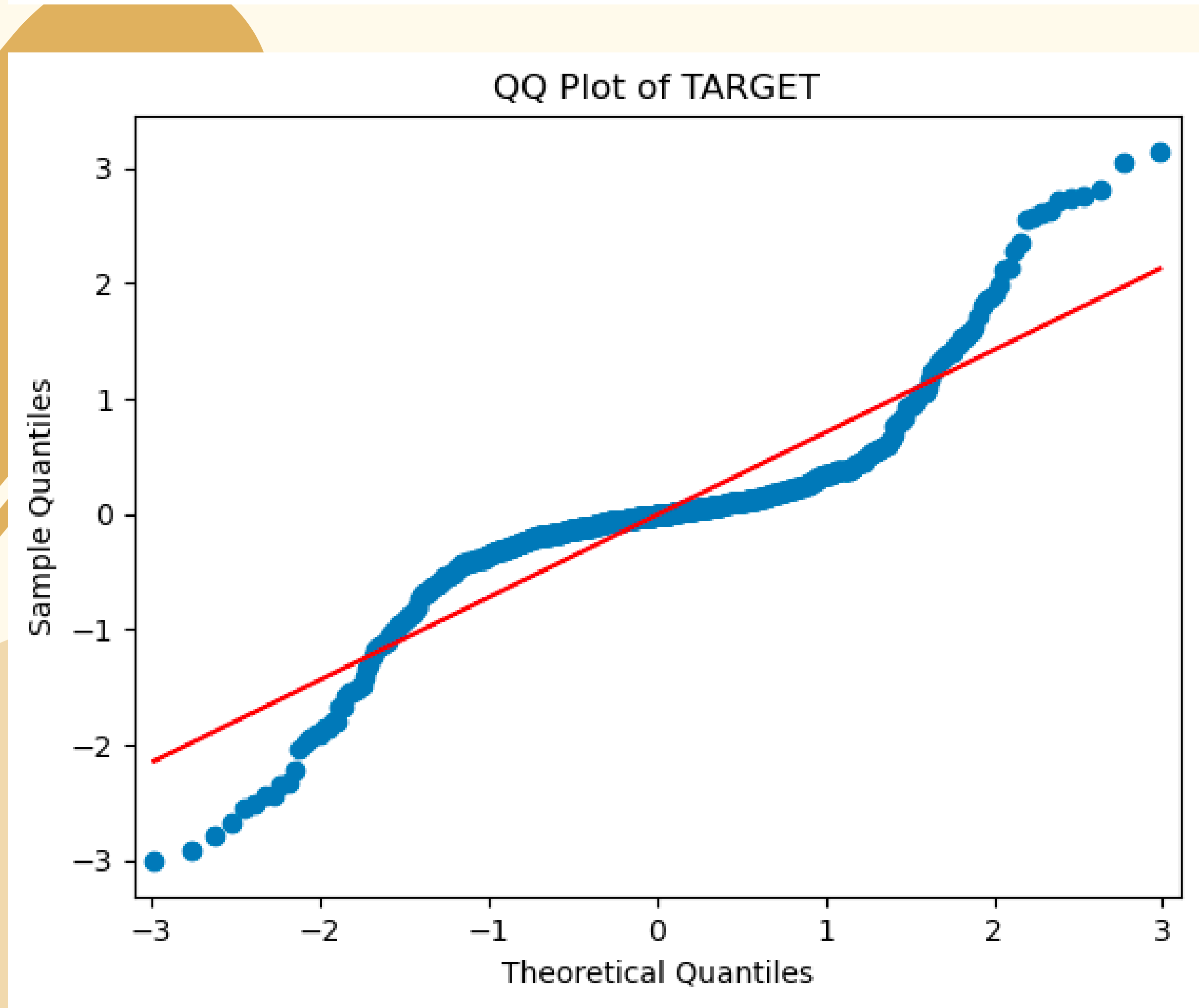
The data is time series data with null value in weather and import & export & exchange imformation. We tried to fill the null value with the average of the value of previous day and the next day. The result shows that filling of weather is valid, others not. Because filling weather makes more sense.

PREPROCESSING

OUTLIER

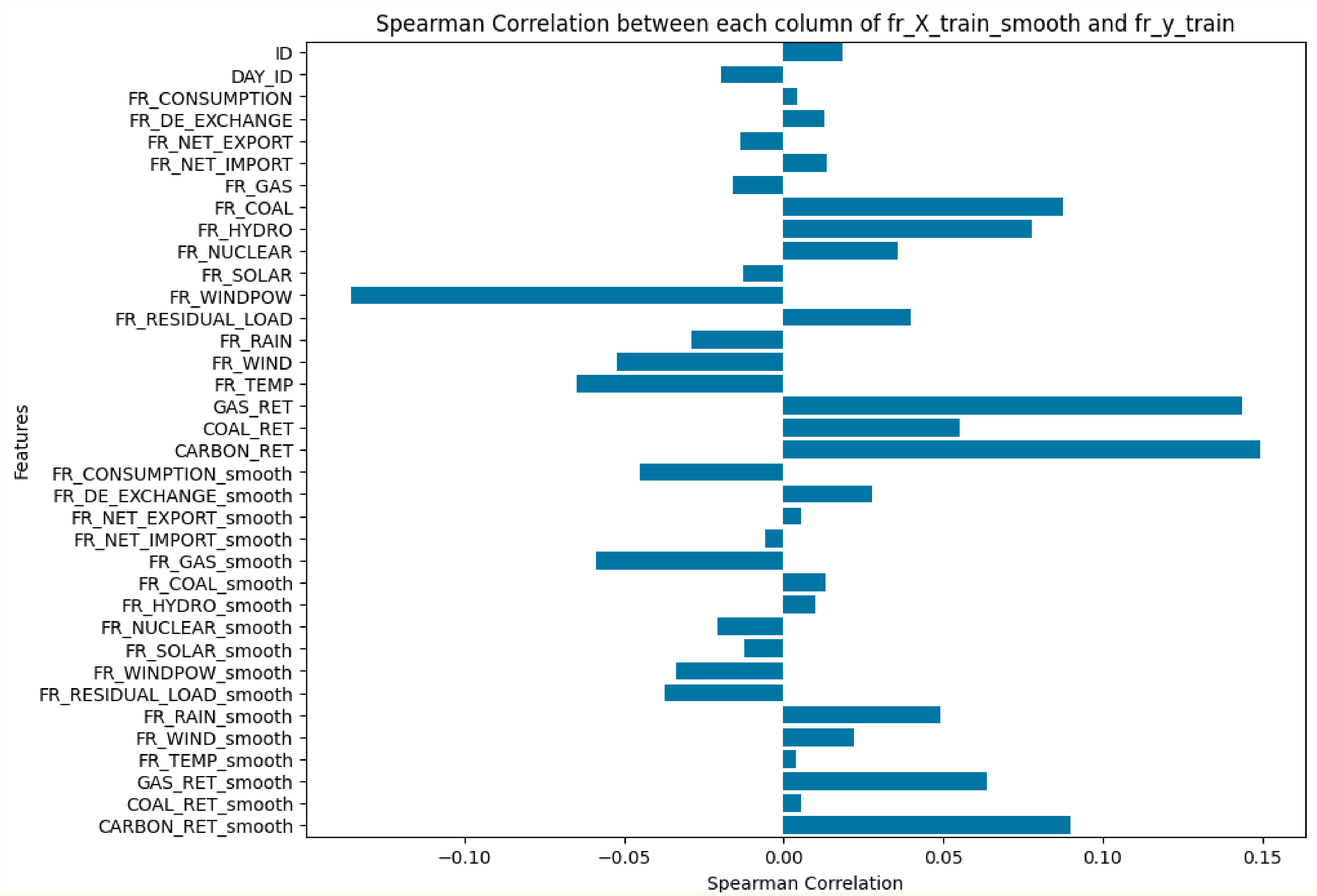
We find that there are many outliers in france dataset, so we delete some data according to Three Sigma Rule. The result shows that removing outliers works well.





PREPROCESSING

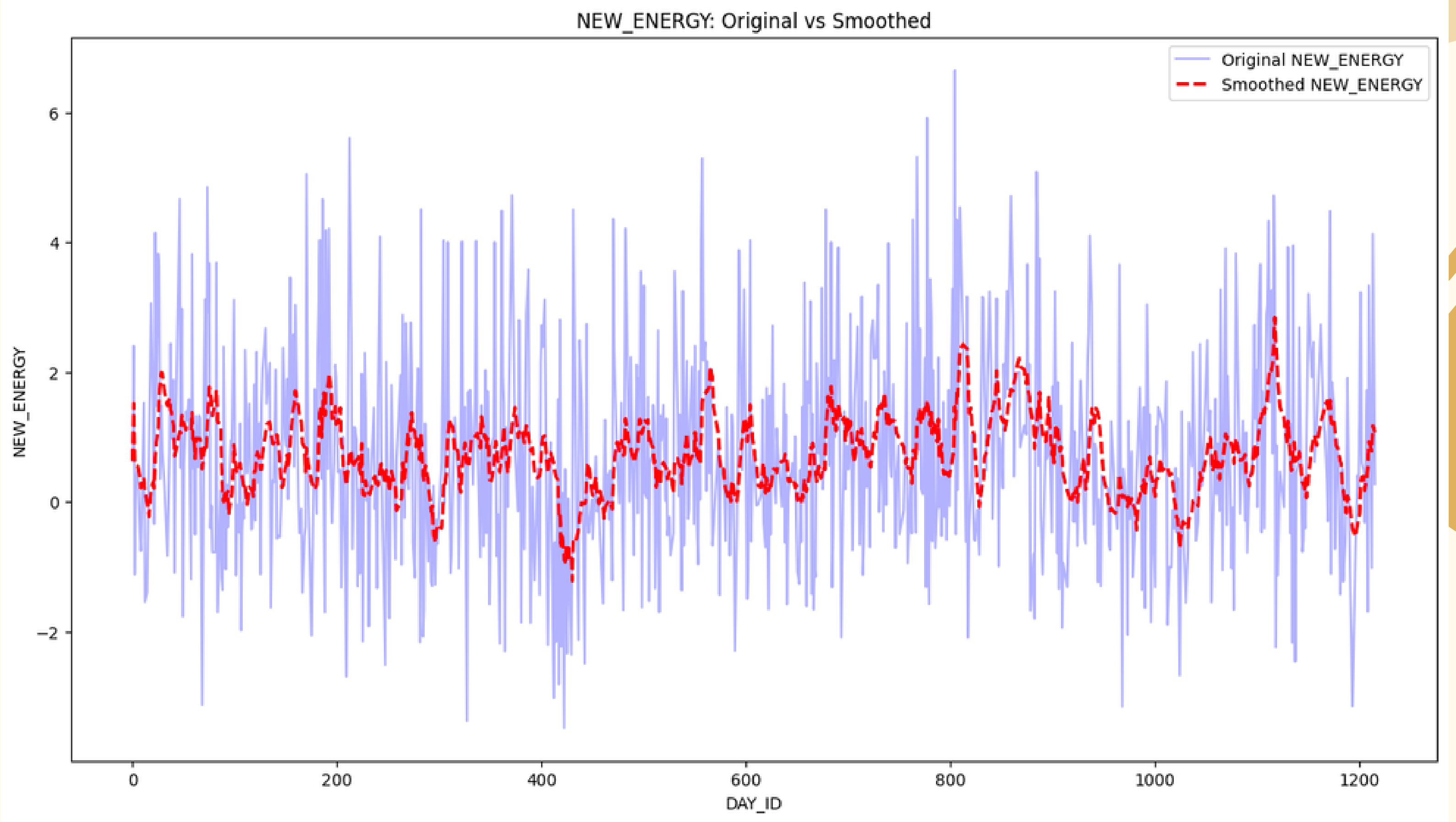
SMOOTHING



*ID and DAY_ID was not smoothed

We smooth the data by creating a rolling windows of 3, 5, and 10 days separately.

We use rolling windows as a parameter for tuning the model performance, and discovered that the model performance decrease slightly with smoothing tehchnique.

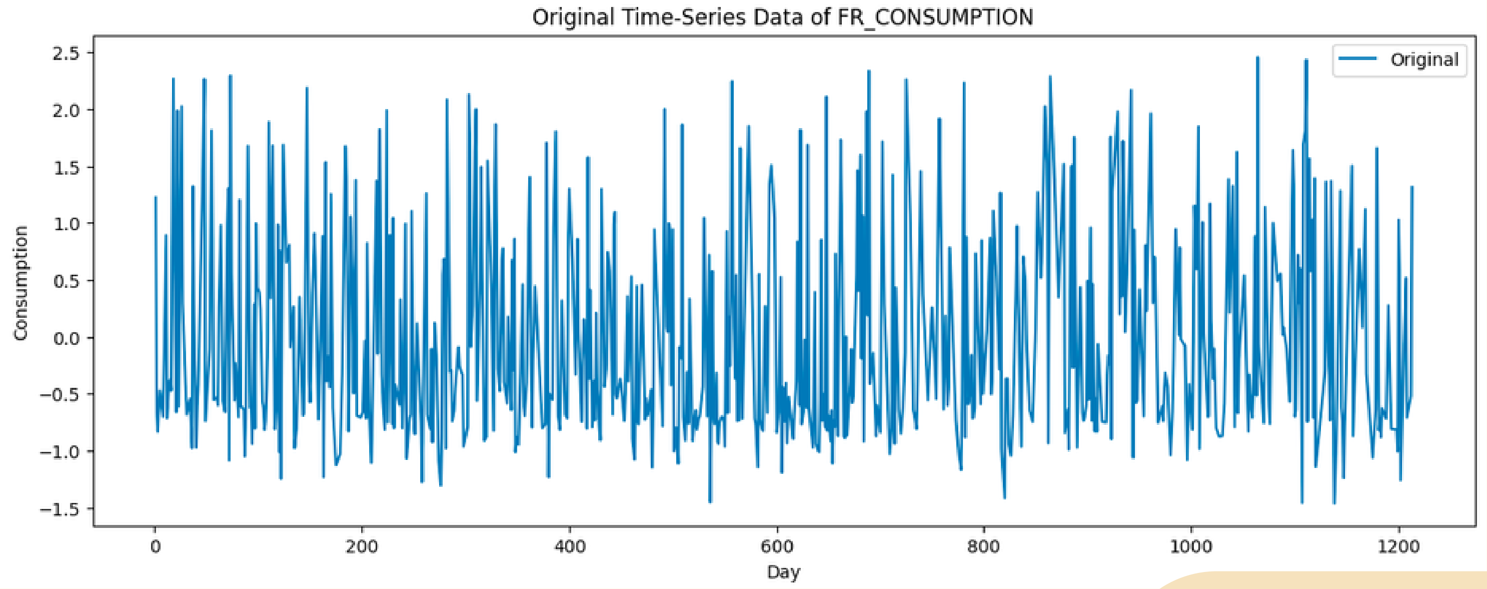


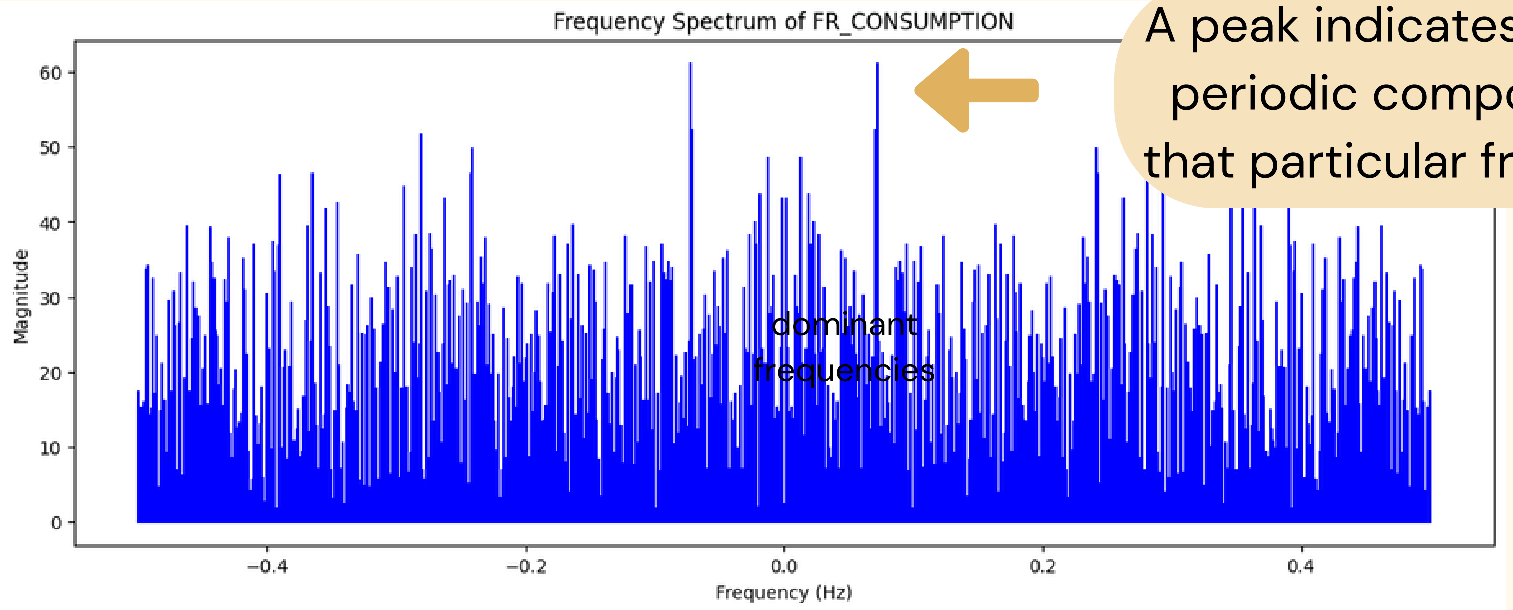
We therefore deduce the followings are potential reasonings for this outcome：

- **Smoothing may lead to information loss**: It often reduces data volatility and noise, potentially removing important predictive signals.
- **Nature of the target variable:** The target exhibits meaningful volatility, thus smoothing such fluctuations could negatively impact predictions.

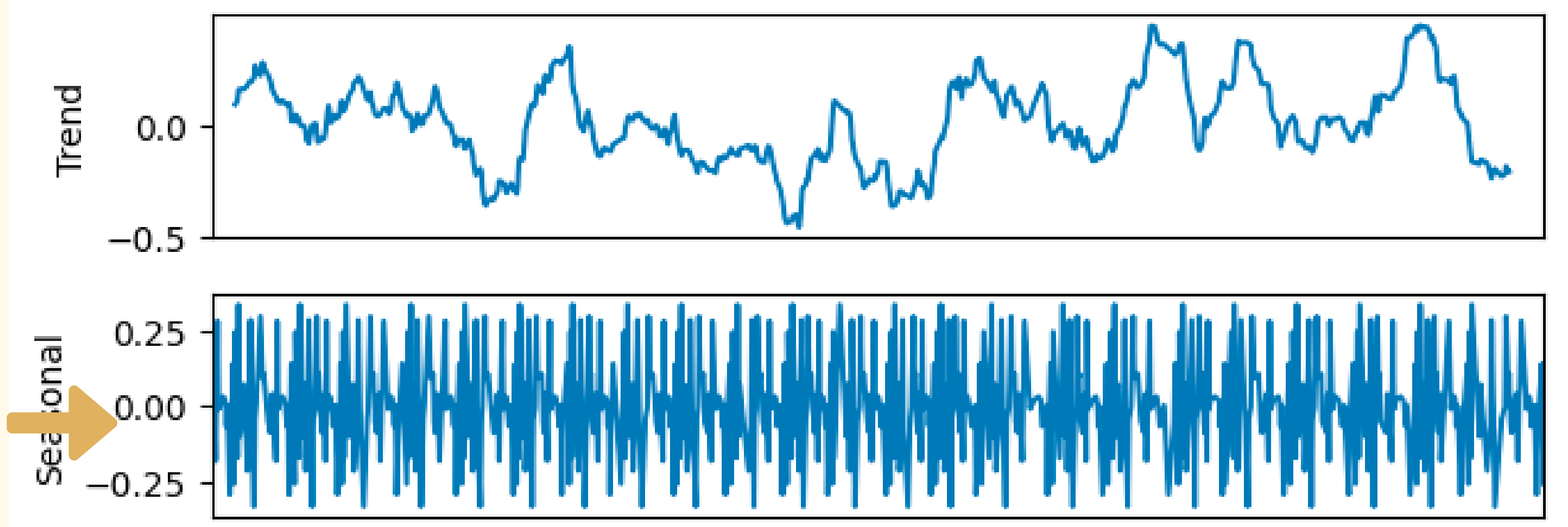
PREPROCESSING

Data Transformation





A peak indicates a strong periodic component at that particular frequency.



due to the limited number of time series data, we set the period as 30 days.

X-axis (Frequency in Hz):how often a pattern repeats per unit of time.
Y-axis (Magnitude): how "strong" or "dominant" each frequency is in the time series data.

Since some features perform better in the original form, we used a greedy algorithm to decide the best features to be transformed in order to optimize the model performance.

Meaningful volatility in target variable

Meaningful volatility in independent variable

The nature of **time series data**

Detecting meaningful frequency with **Fourier Transform**

FEATURE ENGINEERING

- **Sum up all renewable and non-renewable energies**

  We sum up all the renewable energies including HYDRO, NUCLEAR, SOLAR, WINDPOW as a new variable NEW_ENERGY. For German, add DE_RESIDUAL_LOAD on it. We also sum up all the non-renewable energies including GAS, COAL as a new variable OLD_ENERGY.

- **Get weekday using DAY_ID**

  We divide DAY_ID by 7 to get the reminder, then plus 1 as the new variable WEEKDAY. Because we think the price of different weekdays may differ.

- **PCA**

  We use PCA for dimensionality reduction and feature extraction for DE model. And the result shows that

FEATURE ENGINEERING

- **Germany's electricity supply market**

  DE_CONSUMPTION/DE_NET_IMPORT reflects the relationship between domestic electricity demand and imported electricity

We also tried:

- coal price (COAL_RET) * coal production (x_COAL)
- gas price (GAS_RET) * gas production (x_GAS)
- FR_NEW_LOAD = FR_CONSUMPTION - FR_RESIDUAL_LOAD
- delete FR-DE-EXCHANGE, DE_FR_EXCHANGE
- DE_CONSUMPTION/DE_FR_EXCHANGE

MODELS & HYPERPARAMETER TUNING* - FR

Random Forest

- EL method (averaging)
- Noticeable uplift by avoid overfitting, which is realized by decreasing the estimators and min sample leaf

Best Parameters: {'bootstrap': True, 'max_depth': 10, 'max_features': 'sqrt', 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 50}
Best Score: 0.24430840361316836

Adaboost

- EL method (iterative)
- barely no uplift through multiple trials of different parameter combinations

Best Parameters: {'learning_rate': 1.0, 'loss': 'linear', 'n_estimators': 100}
Best Score: 0.1533273256206468

KNN

- Simple and instance-based
- Significant uplift from adjusting k and leaf size considering the size of the dataset

Best Parameters: {'algorithm': 'auto', 'leaf_size': 10, 'metric': 'minkowski', 'n_neighbors': 3, 'p': 1, 'weights': 'uniform'}
Best Score: 0.2006783833453384

*: Since the `GridSearchCV` function provided by sklearn package does not contain a scorer that suits the evaluation matrix given, we used `make_scorer` and `spearmanr` (from scipy package) to form over own evaluation matrix.

MODELS & HYPERPARAMETER TUNING - DE

SVR

- Non-linear Modeling
- Effective in High-dimensional spaces and less sensitive to outliers compared to linear regression.
- The difference in performance between C=100 and C=1000 is not significant
- Performance decrease after PCA.

Best Parameters: {kernel='linear', C=100.0, epsilon=0.499}
Best Score: 0.454735017889O8777

Randomforest

- Non-linear Modeling
- High scalable and it can handle large datasets efficiently
- AddDE_CONSUMPTION/DE_NET_IMPORT feature enhance model performance.
- No uplift through mutiple trials of different parameter combination.

Best Parameters: {'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 4, 'min_samples_split': 2, 'n_estimators': 200}
Best Score: 0.3340507602862254З

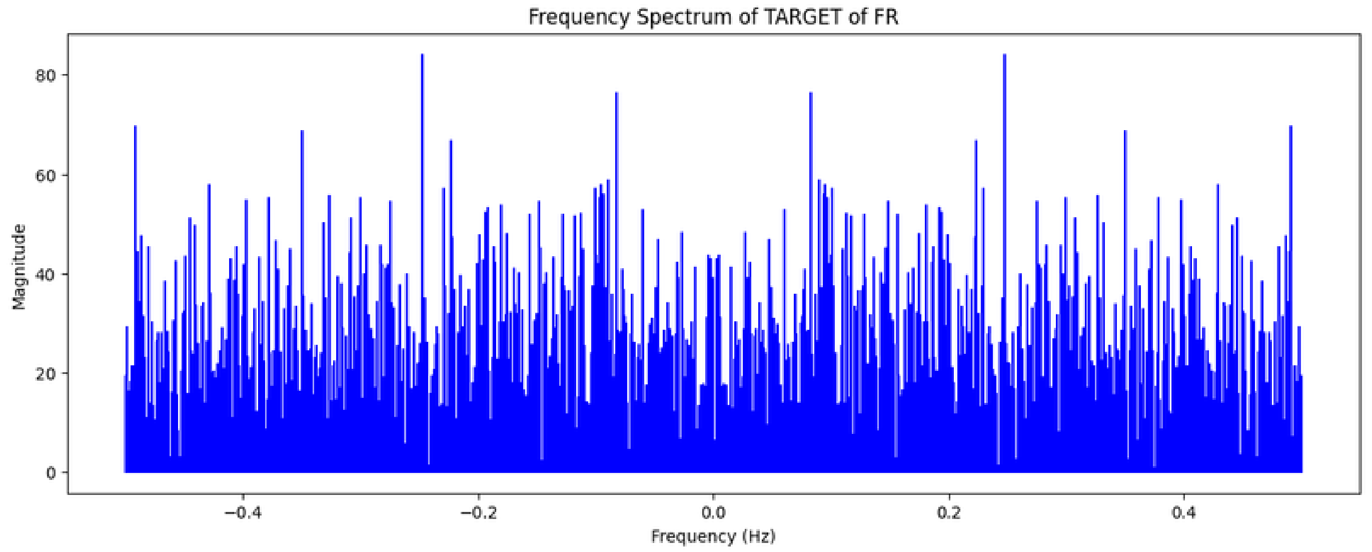
Linear Regression

- Low computational cost
- Offer a balance between simplicity, interpretability, and efficiency
- Performance improves after PCA.

Best Score: 0.522160107334526

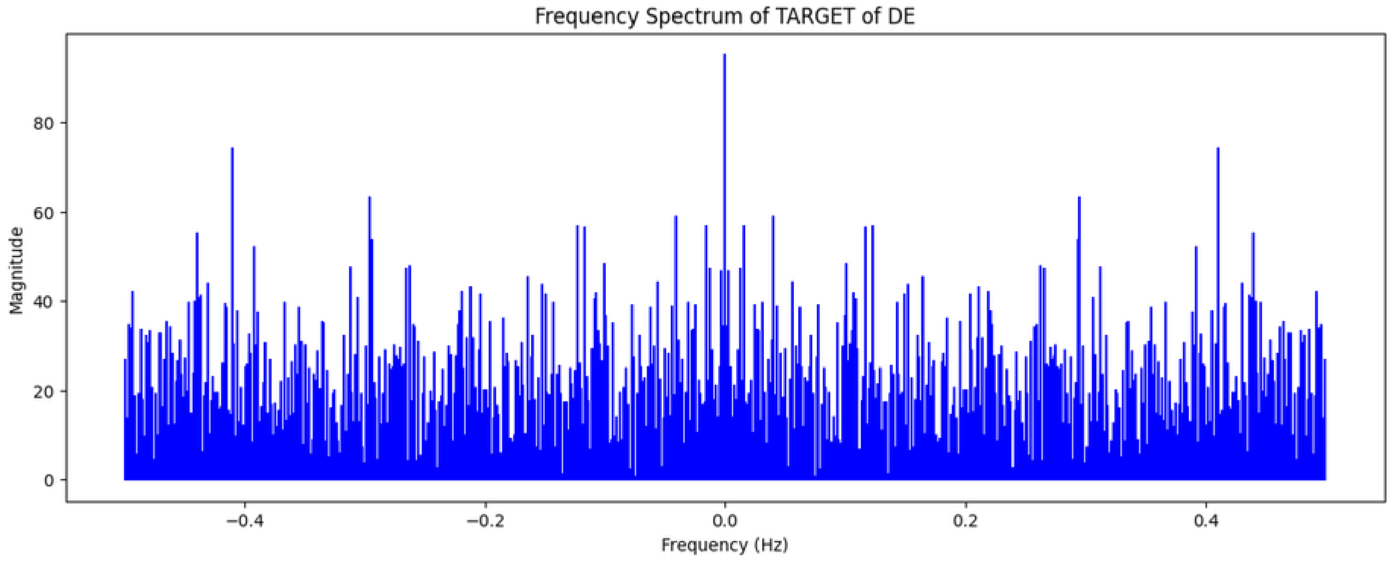
Model Comparison

FR Model



The Fourier transform of the FR model shows a broad range of frequencies with no distinct dominant frequencies or clear patterns. So we can suppose that there is an underlying non-linear pattern. Hence, a flexible model like random forest model will perform better in this senario.

DE Model



The Fourier transform of the DE model shows several distinct dominant frequencies.

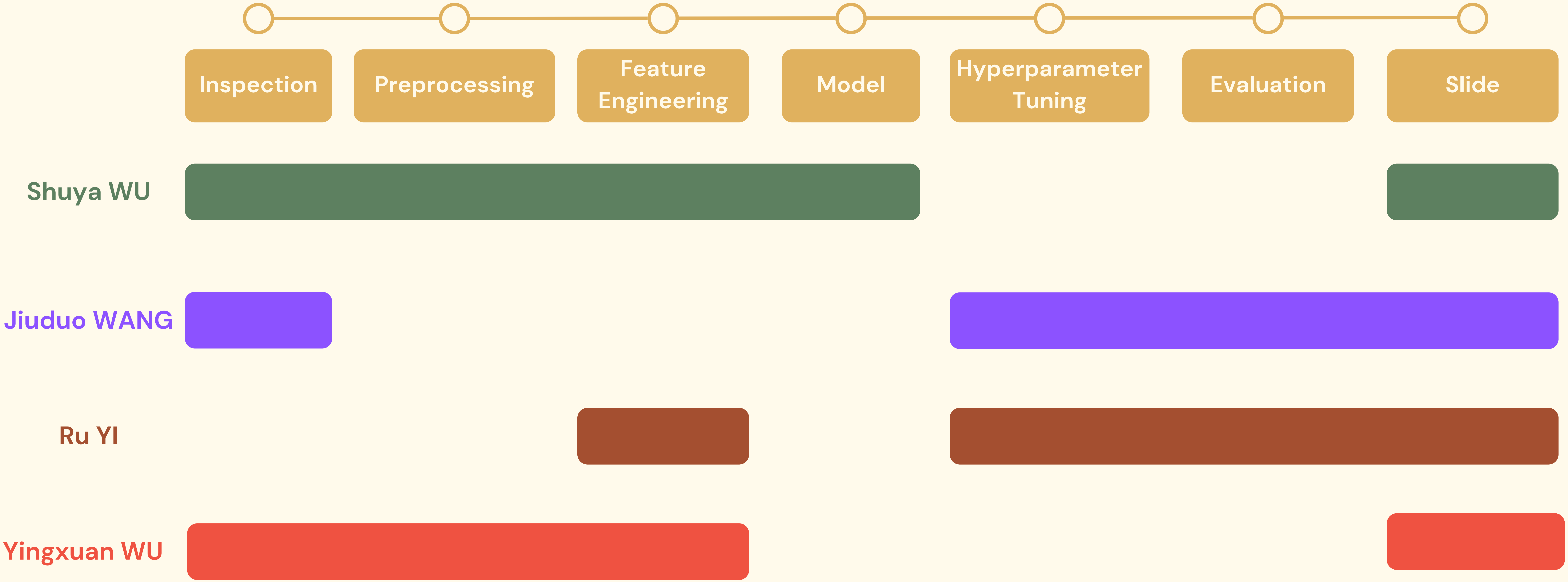
By isolating these strong periodic components and transforming them as linear features, we can capture the underlying linear pattern in the DE section of the dataset.

EVALUATION

Combing both of the model from FR and DE with best performance, we can gave predictions separately on the FR data and DE data on the test set. Return all the predicted values as target, we got our evaluation of the models:

The best score (spearman scorer) we get from Random Forest on France dataset was 0.2443, and the best score on Germany dataset is 0.5221. Applying the models to the test set, we get the result of score 0.2207, where the missing values in test set can have an influence.

PROJECT GANTT



	Inspection	Preprocessing	Feature Engineering	Model	Hyperparameter Tuning	Evaluation	Slide
Shuya WU	🟩	🟩	🟩	🟩			🟩
Jiuduo WANG	🟪				🟪	🟪	🟪
Ru YI			🟫		🟫	🟫	🟫
Yingxuan WU	🟥	🟥	🟥				🟥

THANKS!