Ensemble Learning Project

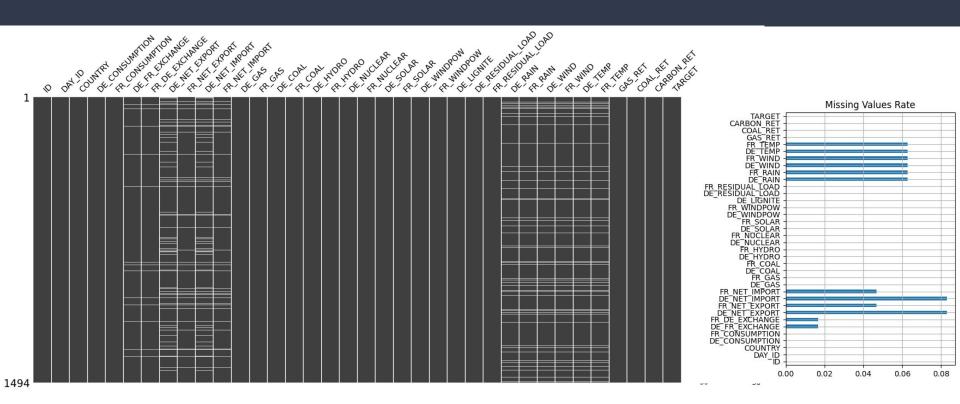
G3-T7

Team Members:

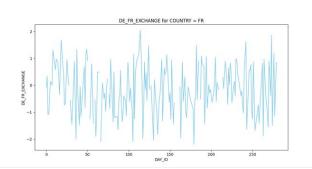
Irene Sunny, Wenjing Zhao

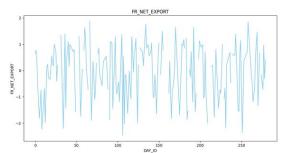
Hongyang Ye, Zheng Wan

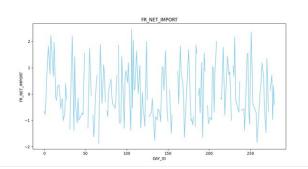
Missing values in features

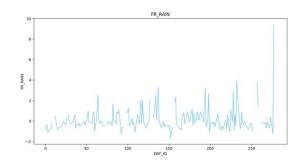


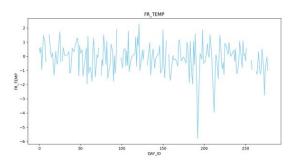
Missing values in features

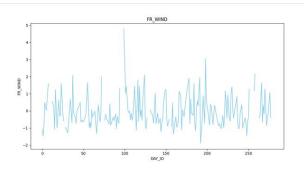










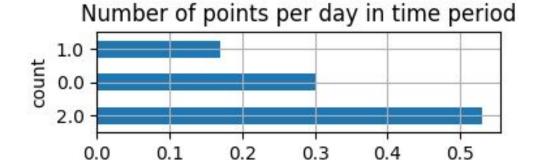


Missing values in sequential data (DAY_ID)

```
# For 208 days we have only information
of FR,no DE

# The occurrences of 'DE' and 'FR' in the
'COUNTRY' column.

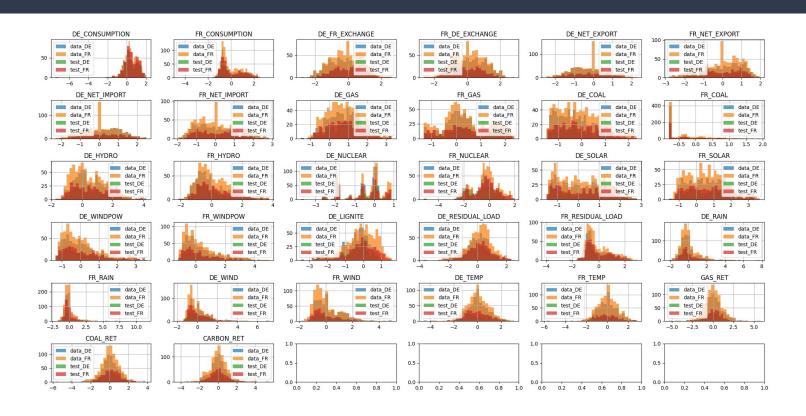
(643, 851)
```



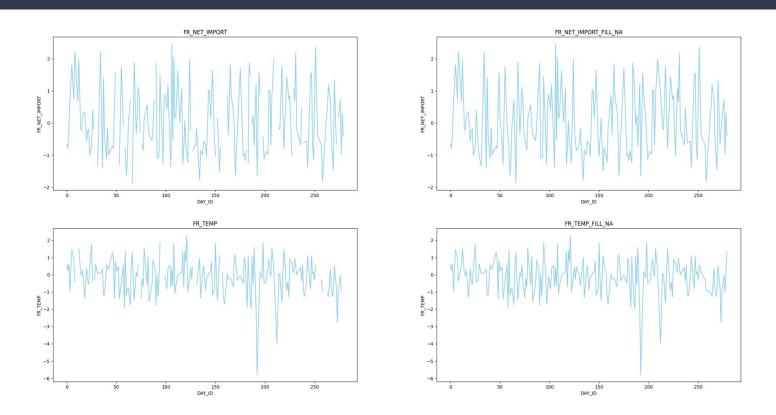
High correlations in features

Correlation Matrix 1.00 - 0.75 FR_NET_EXPORT -0.1-20.2-0.6-0.6-0.2-0.0-2-0.0-2-0.0-2-0.0-2-0.1-20.2-0.1-20.0-1 - 0.50 DE GAS 0.28.38.03.040.08.240.25.20.25.00 160.39.20.09.400.160.140.240.250.480.20.360.70.446.1-0.040.30.28.00.18.050.08.09 DE COAL 0.46.48.39.39.15.36.15.36.39.54.00.60.15.20.10.20.30.25.25.28.10.70.60.50.08.09.09.00.05.16.02.00.05 - 0.25 - 0.00 DE SOLAR -0.69.60.20.270.49.030.440.08.26.50.380.380.380.350.50.040.480.20.20.60.370.19.010.030.08.140.100.00.00 FR SOLAR -).5 \$.60.16.160.40.10.40.120.250.46.25.30.30.40.40.60.80.00.36.40.20.20.60.19.26.20.100.10.00.00.01 DE WINDPOW 4.48.36.39.3 70.0 70.0 70.0 48.10.28.0 50.40.0 0.08.2 40.45.3 400.6 60.3 6.70.2 0.2 0.10.40.2 5.10.0 0.09 FR WINDPOW 4.39.30.06.06.39.130.39.130.39.130.20.080.16.04.30.020.00.180.49.40.67.00.29.39.08.20.20.20.40.00.09.09.09.00.00. - -0.25 DE LIGNITE 4.35.35.220.20.06.24.06.24.38.45.70.420.00.10.18.220.26.240.38.25.00.70.420.1-0.040.24.1-0.05.1-0.040.05.06 DE RESIDUAL LOAD 4.22.26.10.12.340.06.380.06.70.410.60.380.13.370.30.250.220.20.70.340.71.000.440.1-0.020.440.20.1-9.10.06.080.11 -0.50- -0.75

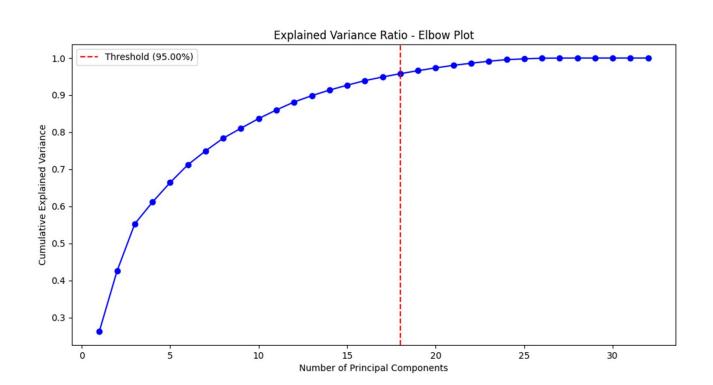
Feature Distribution (Normalization?)



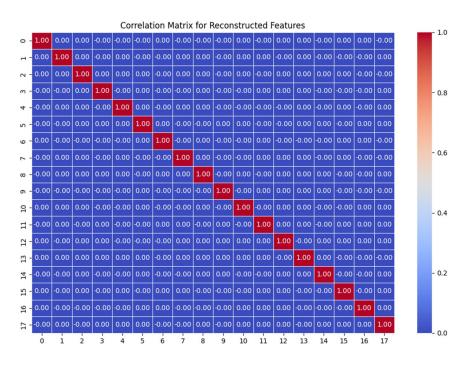
Solution - Missing values



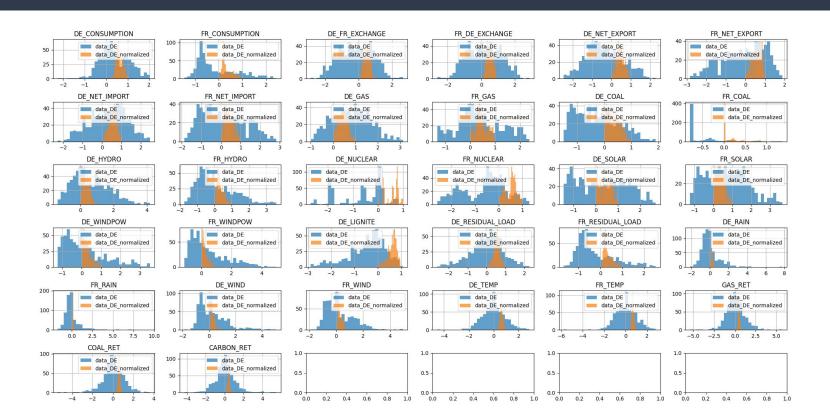
Solution - Multicollinearity



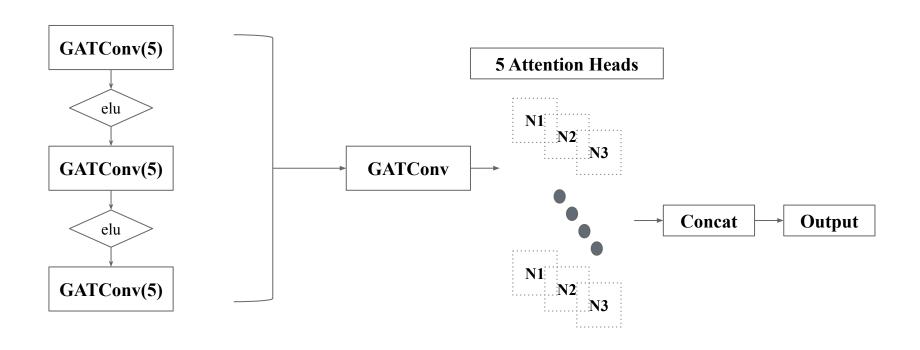
Solution - Multicollinearity



Solution - Normalization



Final Diagram



Summary

Observations

- There are missing values in some features
- There are missing values in days
- There are features that might be highly correlated
- We can't tell if the data is already normalized or not

Solutions

- Replace missing values with appropriate approaches
- Check correlations between features
- Implement PCA on dataset to eliminate the effect of multicollinearity
- Conduct normalization on features