

01

Prediction Modeling

Serival Tree Team

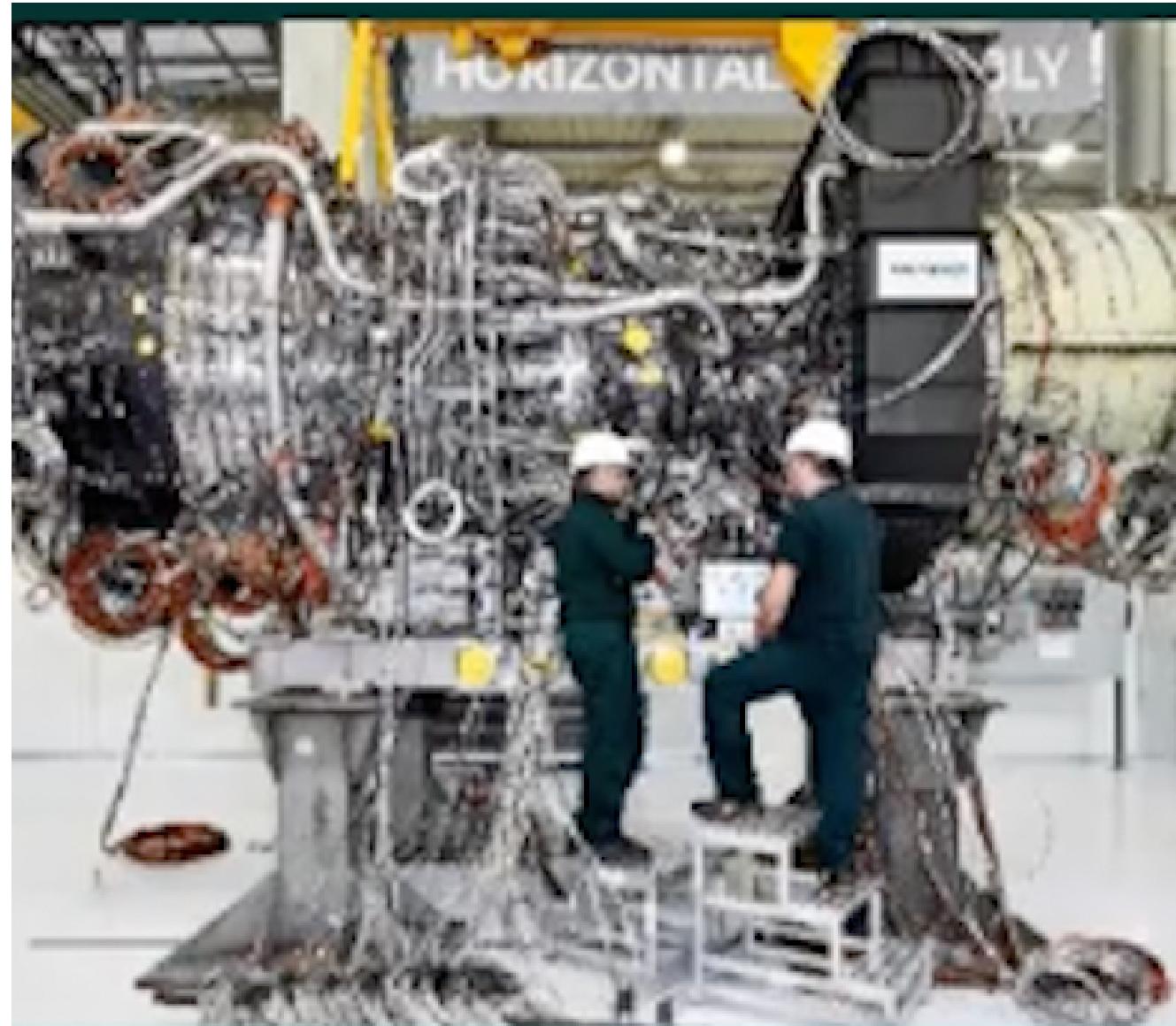
- Ivana Sánchez Olivares
- Sergio Maldonado
- Jorge Alberto Iván Ambriz



Image of a representation of an Artificial Neural Network

What is the Challenge ?

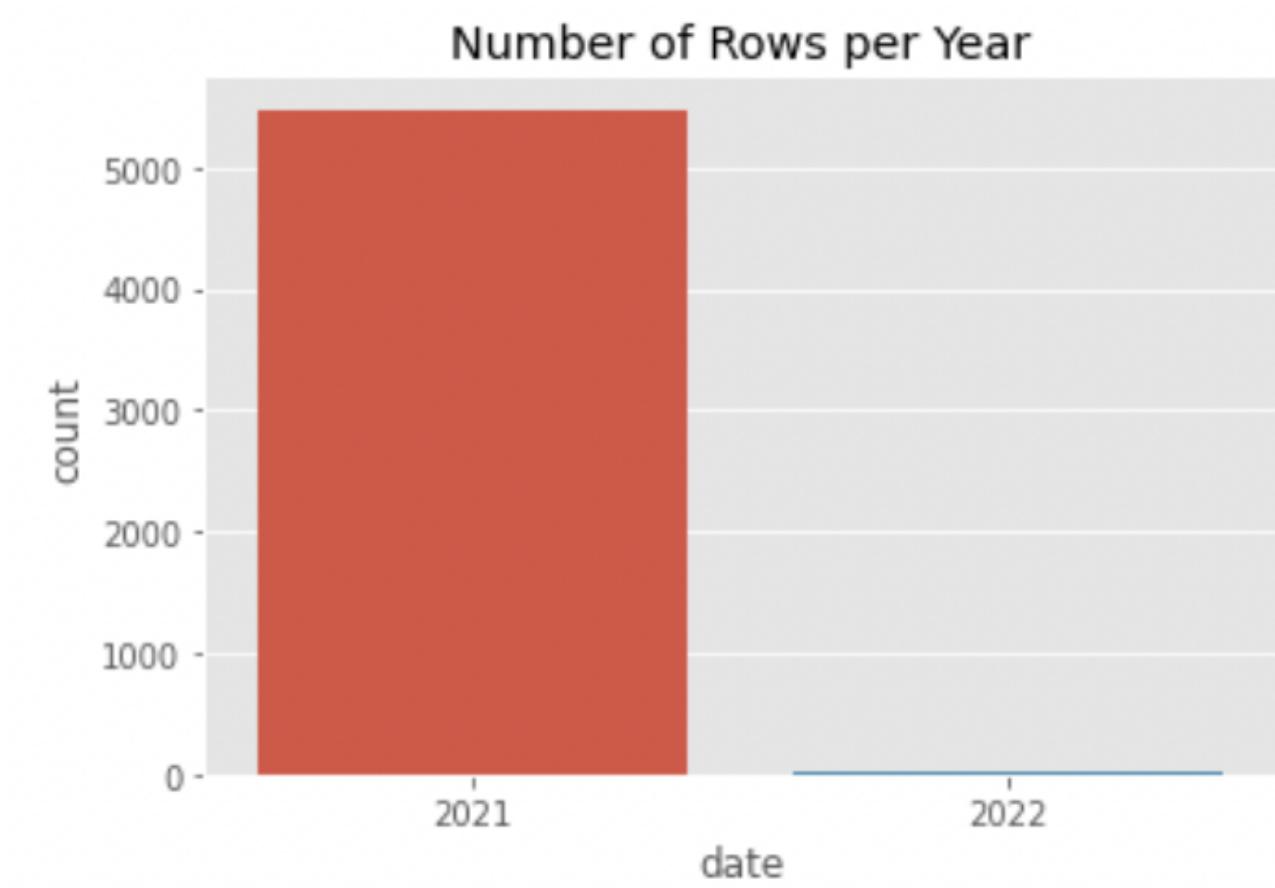
- Worked with synthetic data that simulates the historic behavior of Gas Turbine engines.
- Supporting metadata about the engines and sites is also given.
- **Predict the Low Pressure Turbine (LPT) POWER output for a single engine under certain given conditions**



	POWER	CMP_SPEED	CDP
03	count	4337.000000	5490.000000
	mean	11730.214760	6417.889466
	std	3991.630810	3793.082652
	min	2720.133922	0.000000
	25%	8185.832753	4499.382563
	50%	13473.368134	7823.908803
	75%	14972.032226	9658.852062
	max	17616.227967	10000.000000
			12.390310

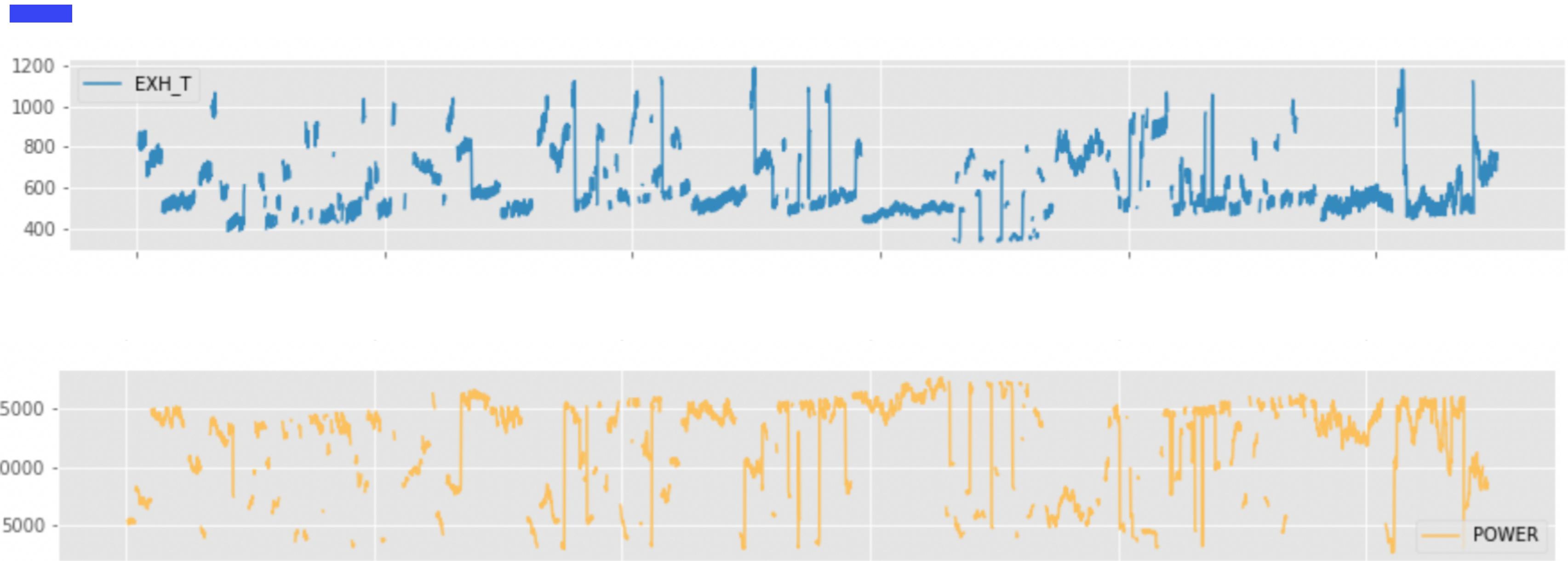
	ELEVATION	HPT_IT	EXH_T
	count	5490.000000	4337.000000
	mean	310.903213	1182.098329
	std	416.506341	118.879552
	min	-29.000000	878.785407
	25%	14.458365	1106.295284
	50%	130.987930	1157.928077
	75%	567.625122	1253.712362
	max	1552.426025	1600.690748
			1188.563234

The train data itself is interesting



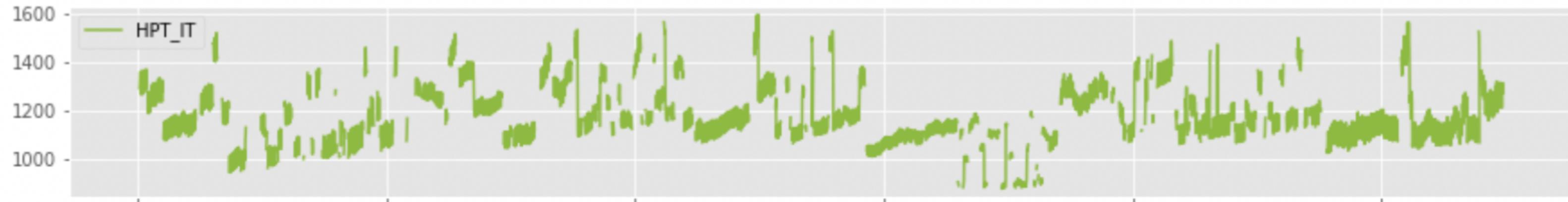
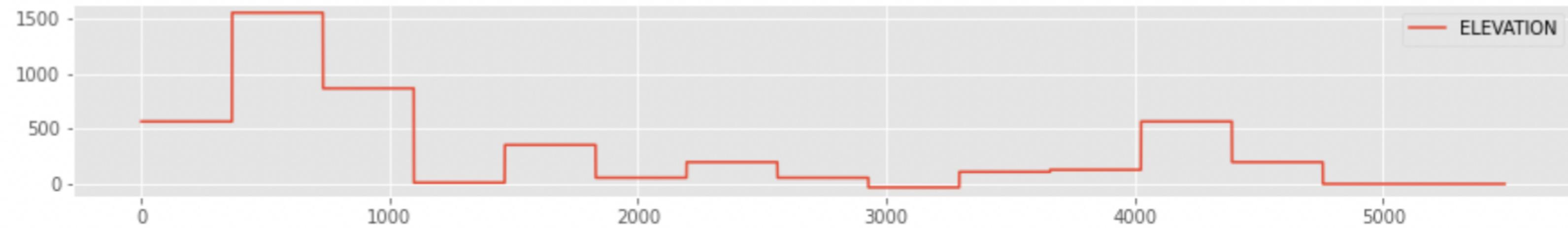
How do variables change ?

04



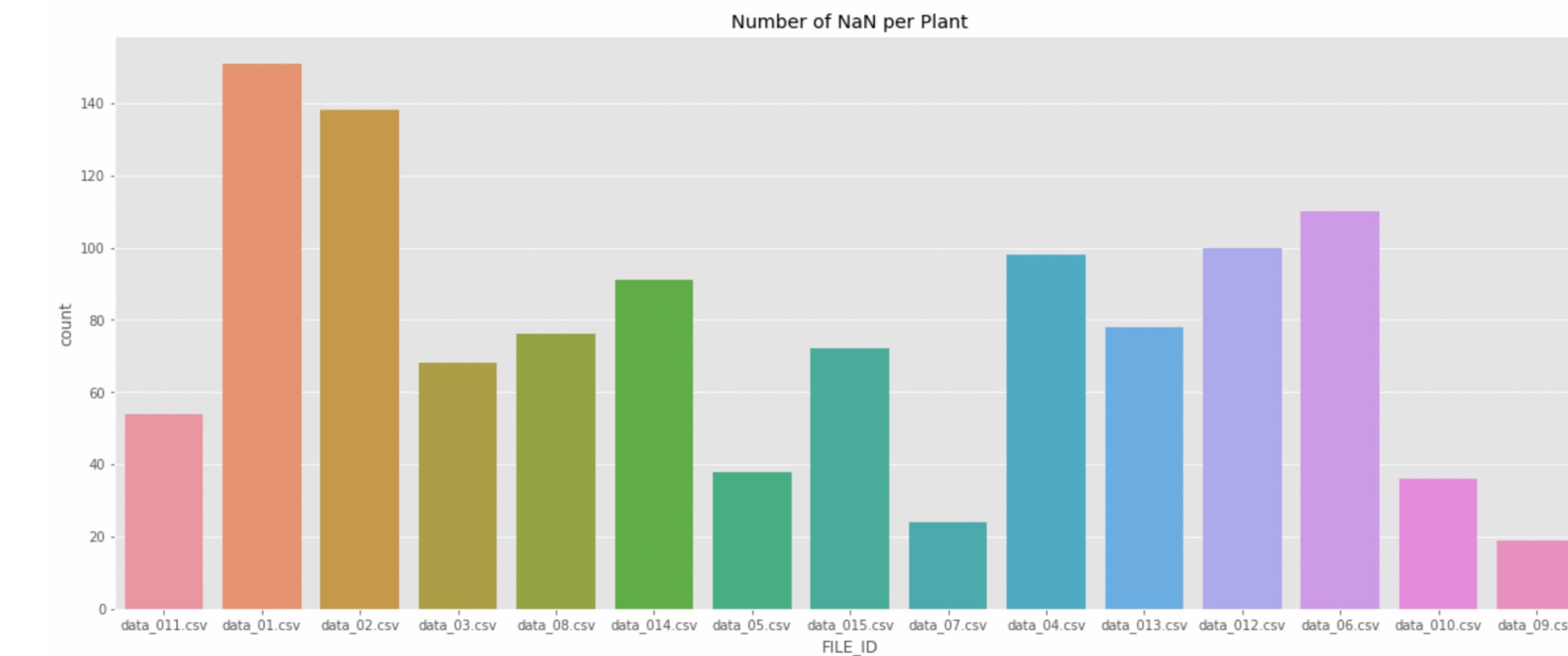
How do variables change ?

05



Is There Any Missing Value ?

06

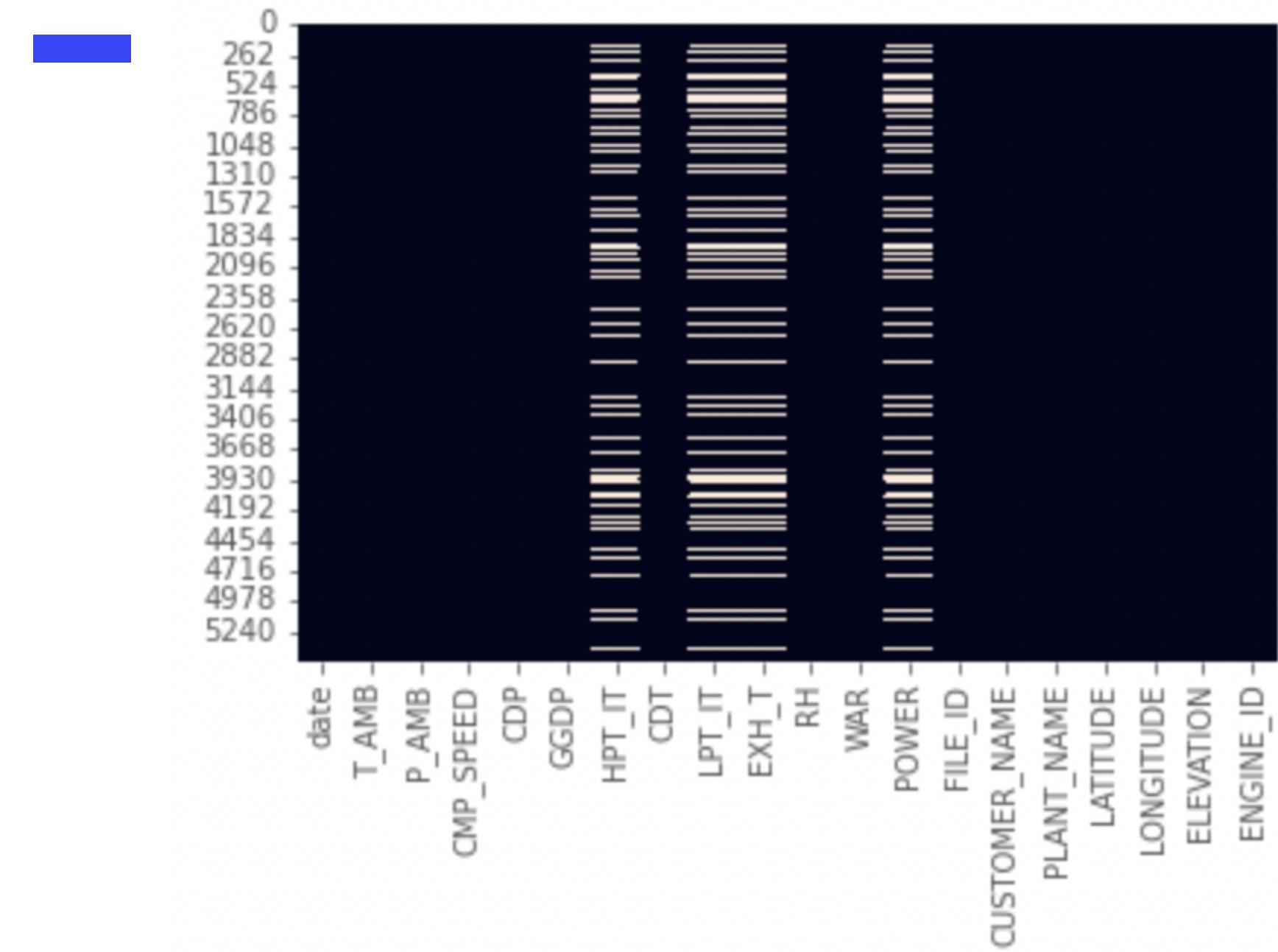


07

What Percentage of them are?

date	0.000000
T_AMB	0.000000
P_AMB	0.000000
CMP_SPEED	0.000000
CDP	0.000000
GGDP	0.000000
HPT_IT	21.001821
CDT	0.000000
LPT_IT	21.001821
EXH_T	21.001821
RH	0.000000
WAR	0.000000
POWER	21.001821
FILE_ID	0.000000
CUSTOMER_NAME	0.000000
PLANT_NAME	0.000000
LATITUDE	0.000000
LONGITUDE	0.000000
ELEVATION	0.000000
ENGINE_ID	0.000000

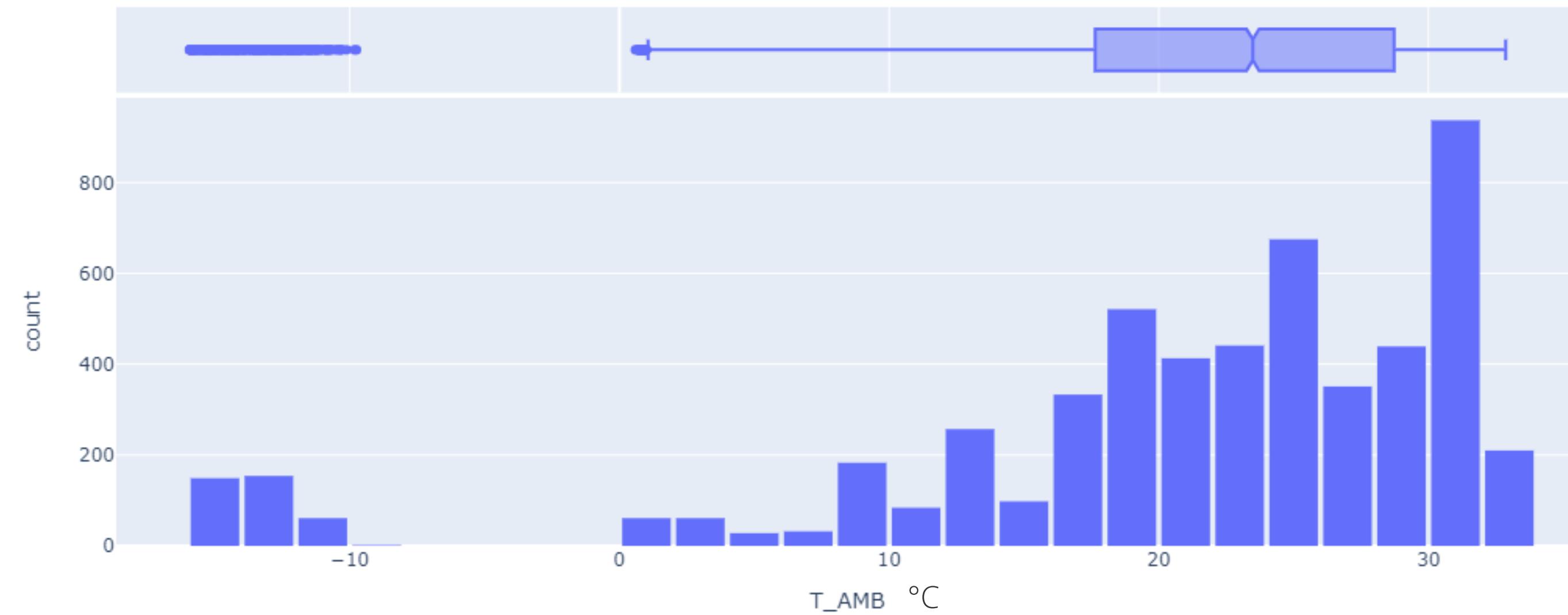
Does NaN values are related ?



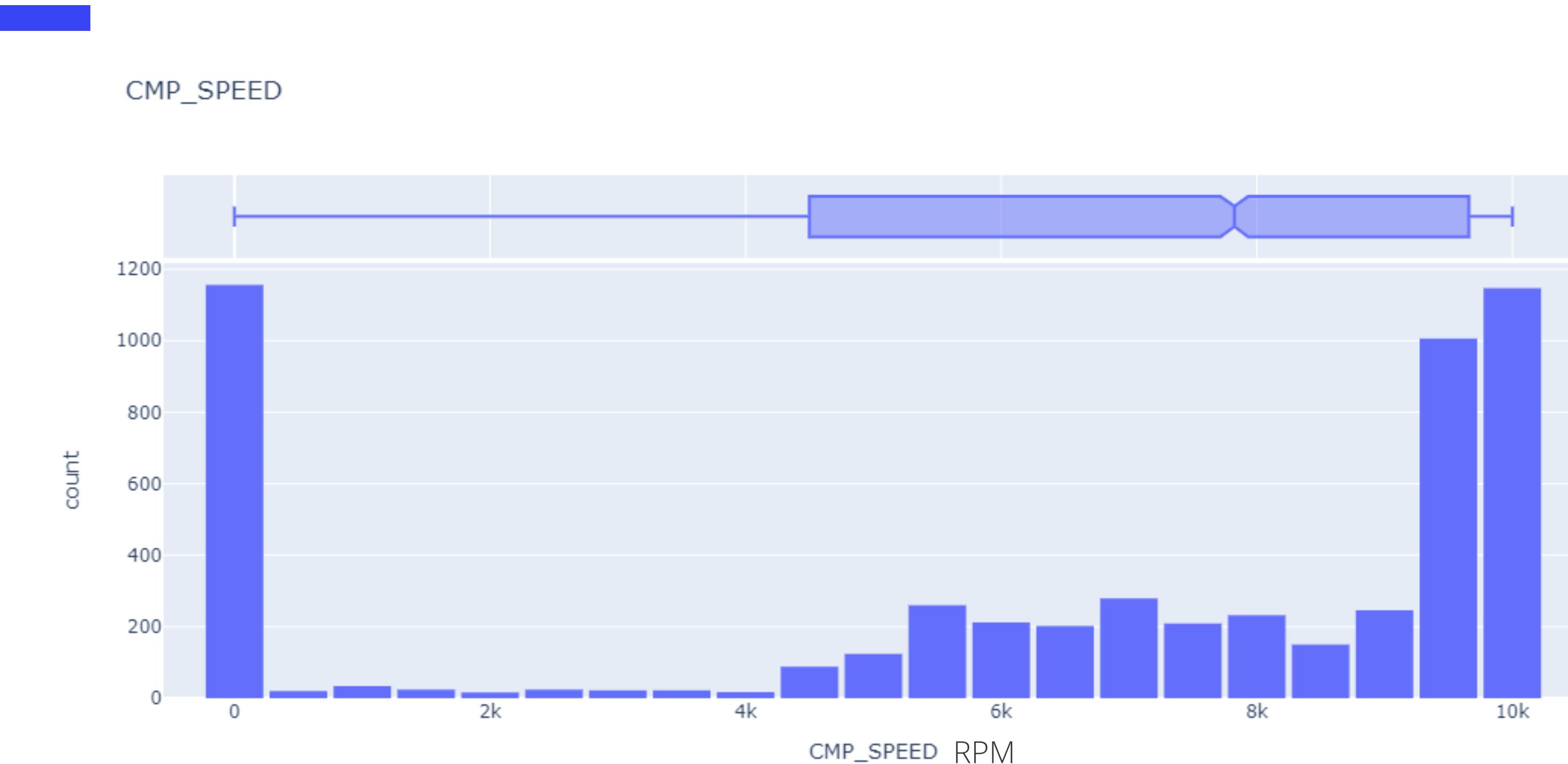
Are there Atypical Values ?



T_AMB

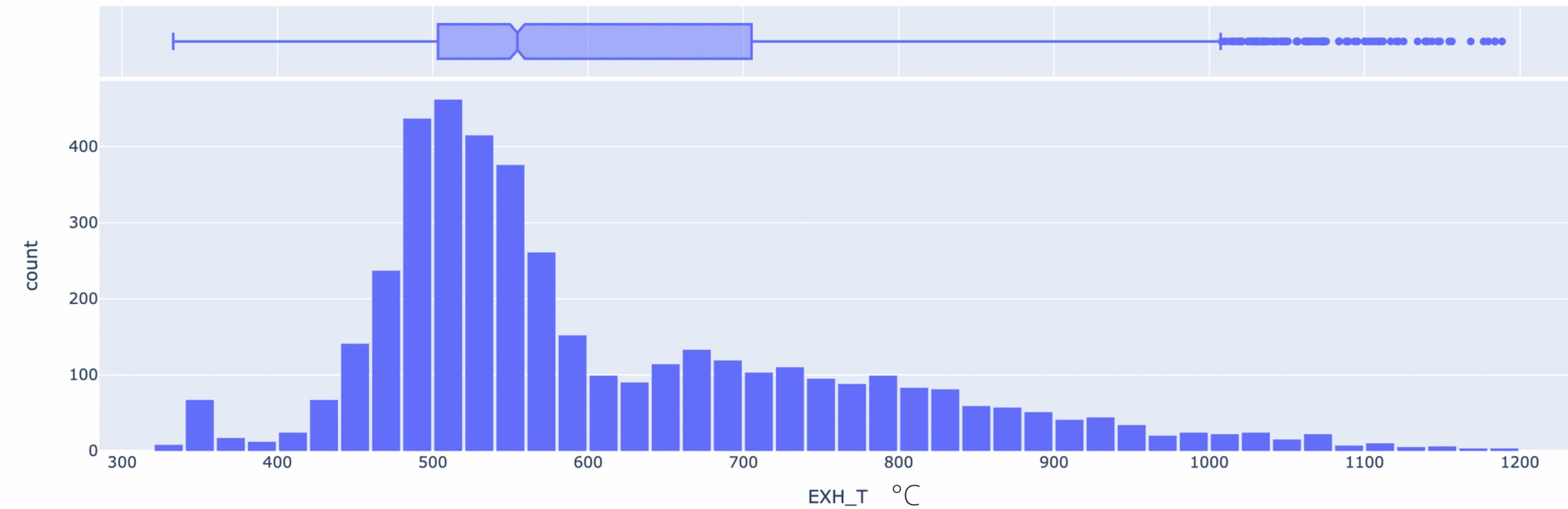


Are there Atypical Values ?



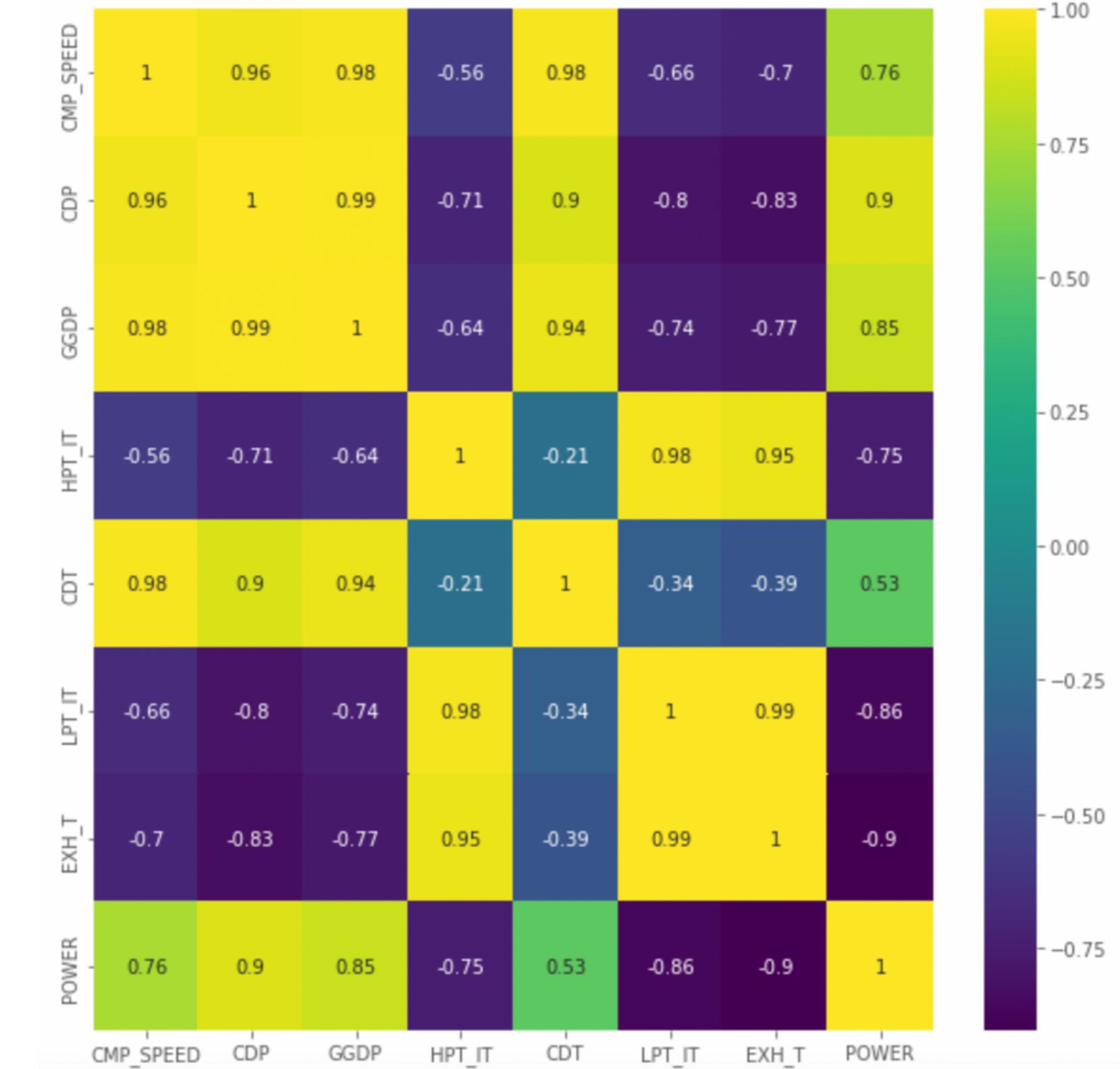
10

Are there Atypical Values ?



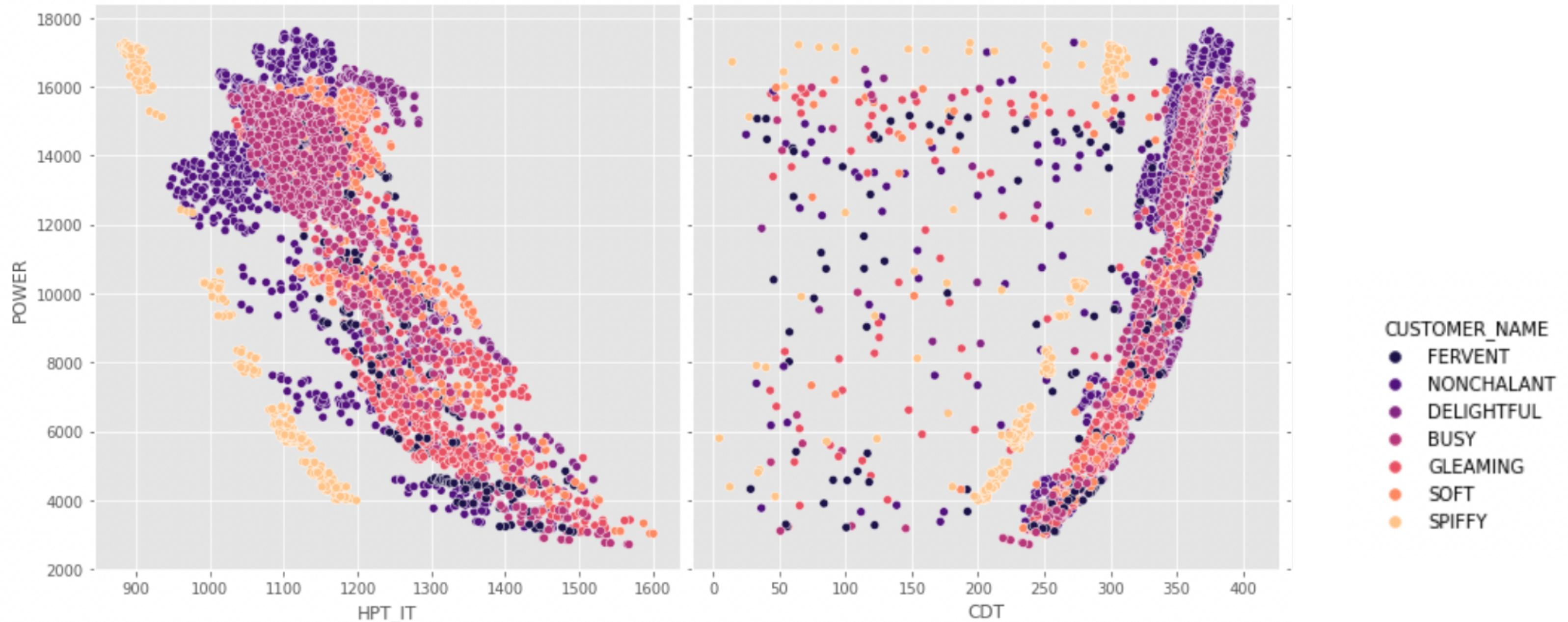
¿How the variables influence on POWER ? The Correlation is here to help.

	T_AMB	-0.004338
	P_AMB	0.155483
	CMP_SPEED	0.758151
	CDP	0.898743
	GGDP	0.846911
	HPT_IT	-0.746706
	CDT	0.525526
	LPT_IT	-0.863357
	EXH_T	-0.903272
	RH	0.017927
	WAR	-0.049835
	POWER	1.000000
	LATITUDE	0.048687
	LONGITUDE	0.034754
	ELEVATION	-0.152488



The Correlation in graphs

12



Model Selection

1

ANN

- Easier to implement
- Good with non-linear shapes
- Experience of the team

2

Random Forest Regression

- High accuracy
- Scales well
- Interpretable
- Easy to use

Data Preprocessing

The variables used are :

- 'CMP_SPEED',
- 'CDP'
- 'GGDP'
- 'HPT_IT'
- 'CDT'
- 'LPT_IT'
- 'EXH_T'

Variable selection is based on EDA.

We use 'atypical values' and all the datas provided in the training

Standardize the data:

- MinMaxScaler for Random Forest and ANN
- Normalization for ANN

Don't use

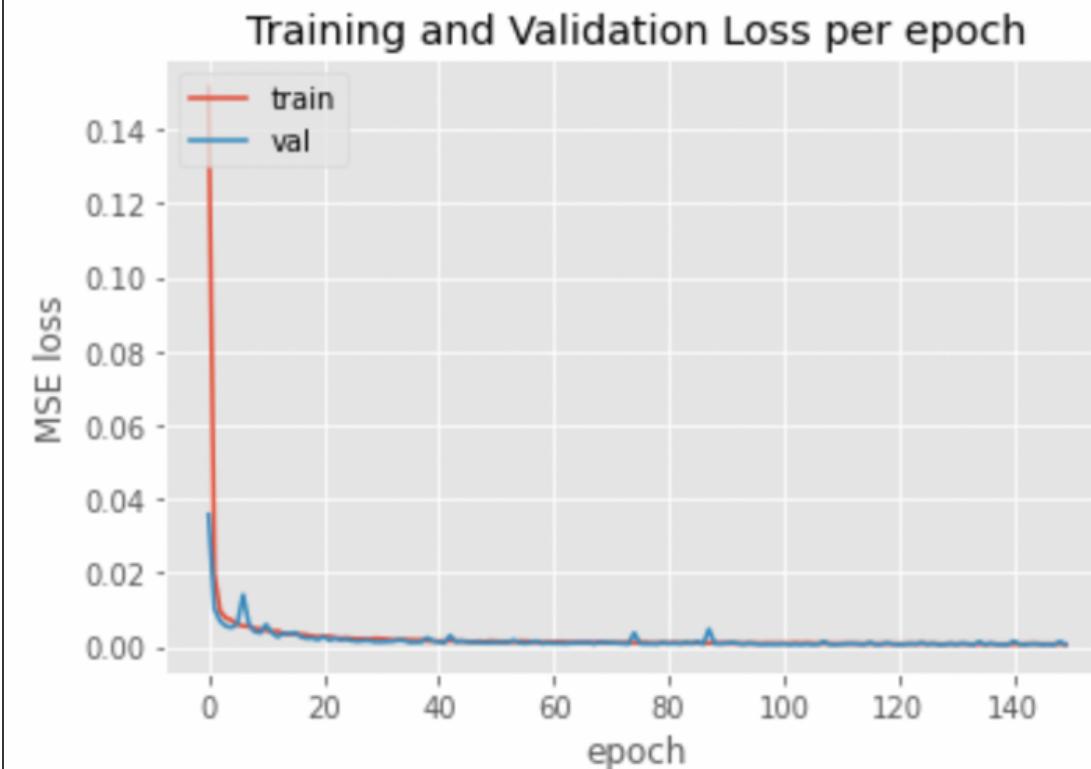
Nan for training because it gives better results in our models

ANN with pytorch

It is an Artificial Neural Network with :

- 3 layers .
- Sigmoid and ELU activation functions.
- SGD optimizer

Let's look how well is the training

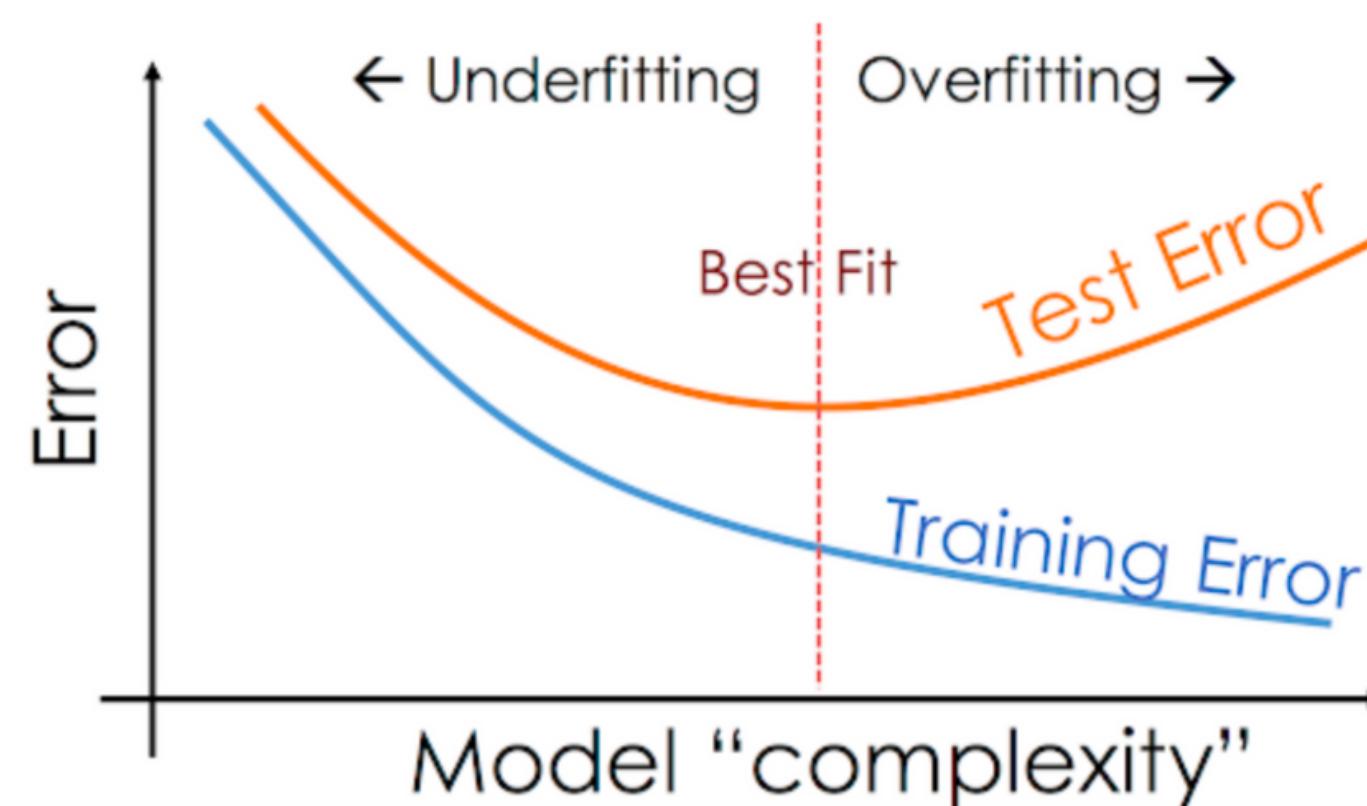


What about
Kaggle Score ?

Best
Score: 1209.98

Worst
Score: 1561.72

Random Forest Regression

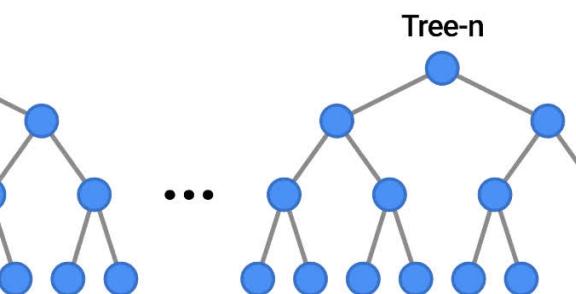
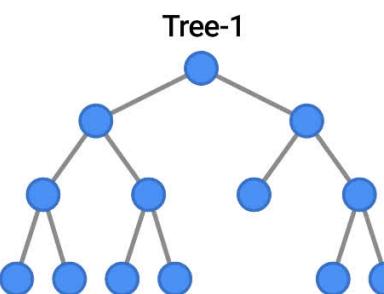
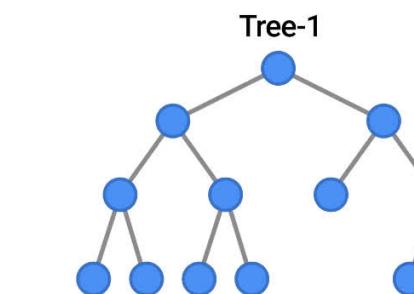


1

2

3

EXAMPLES



Base

RMSE: 328

Kaggle Score: 809.9

Best

RMSE: 341

Kaggle Score: 782.6

Compressor discharge pressure



Exhaust temperature

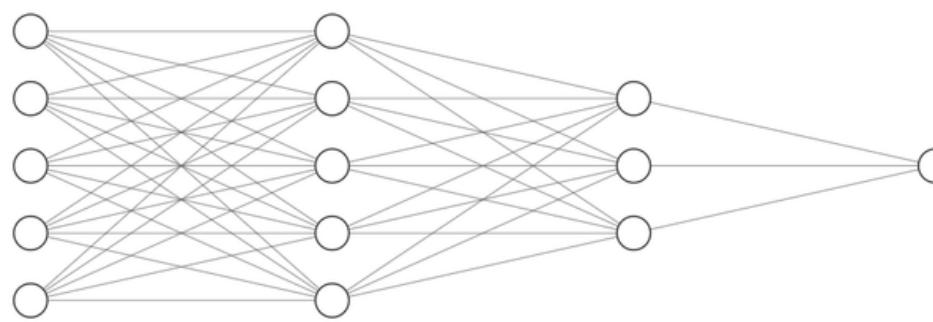


Compressor speed

Gas generator discharge pressure

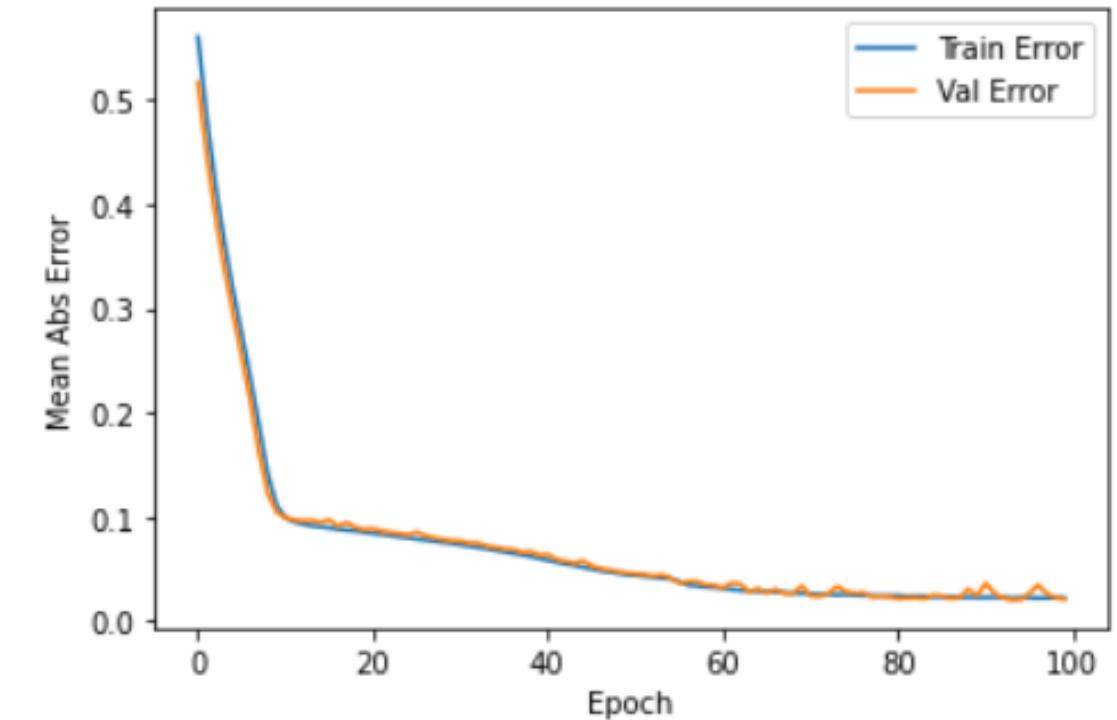
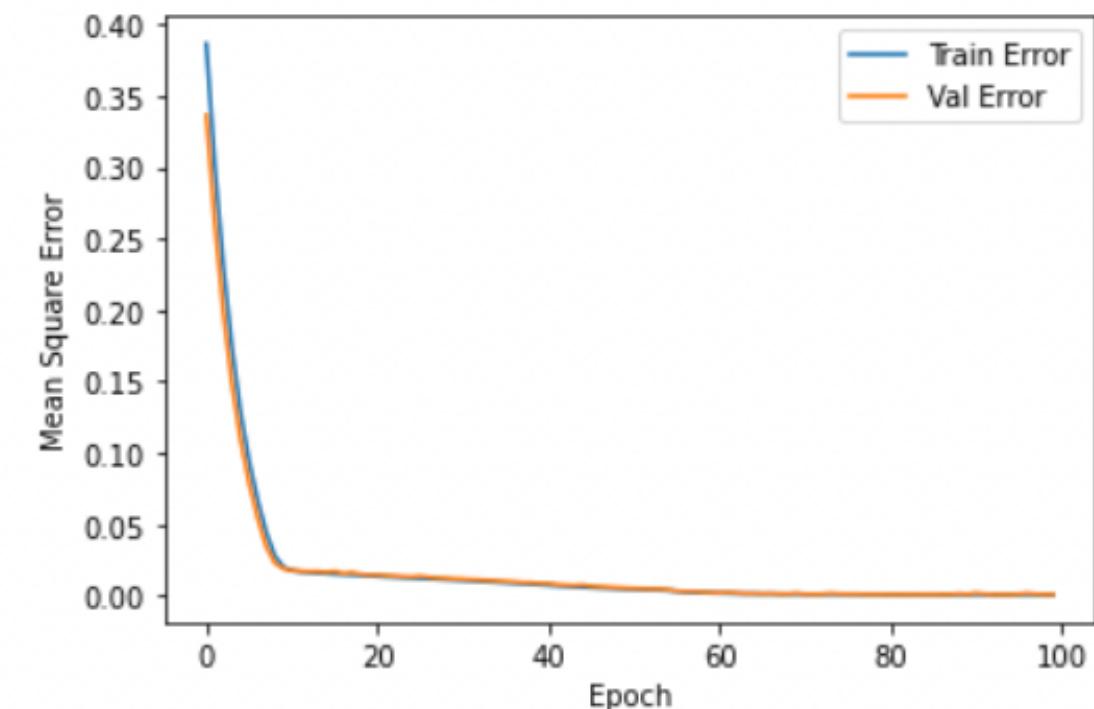
ANN with Tensorflow and Keras

It is an Artificial Neural Network with :

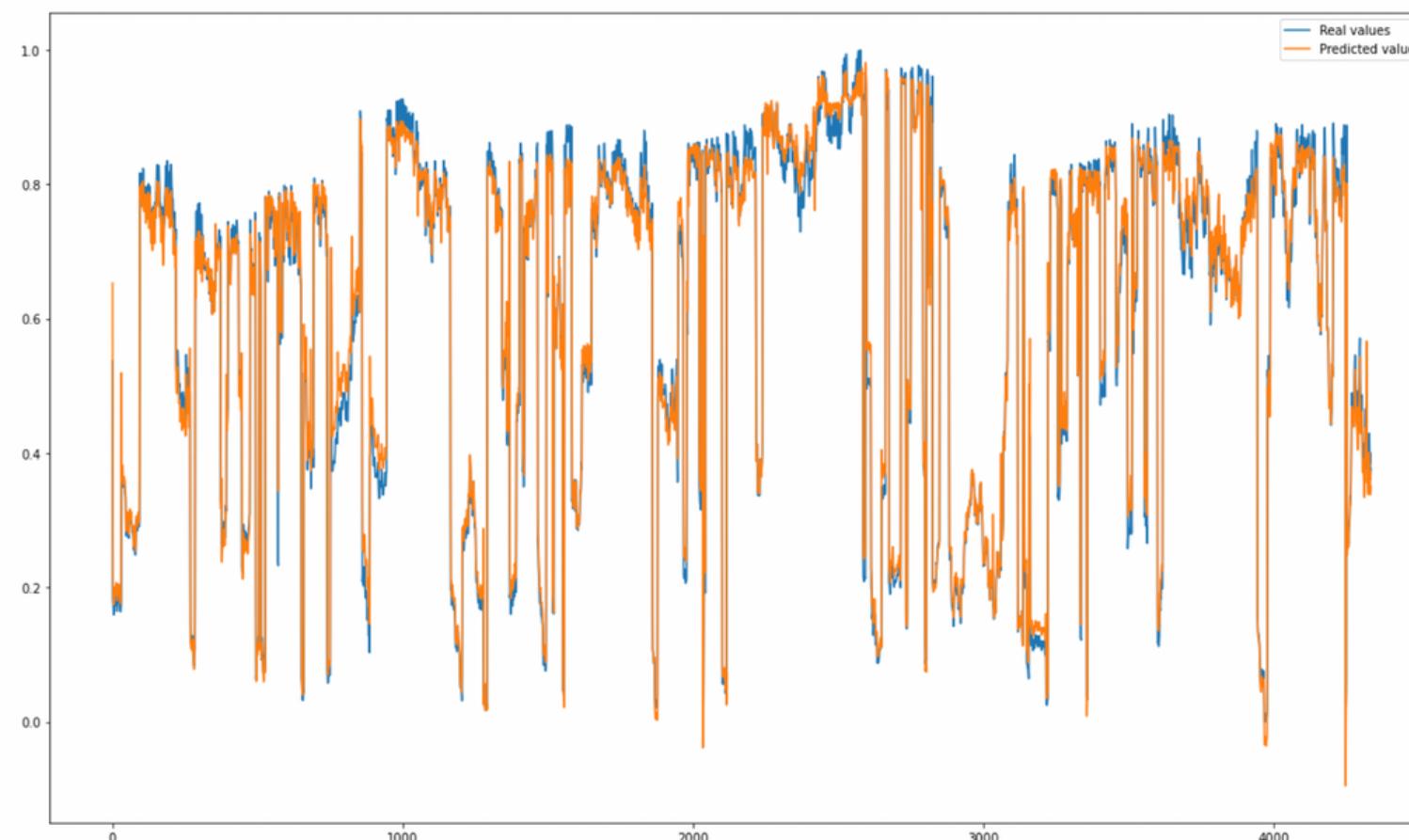


- RMSprop optimizer
- relu activation function

Let's look how well is the training



Real vs prediction in the training data



What about Kaggle scores?



Best

Score: **362.42**

Worst

Score: 1141.85

Average

Score: 847.85



19

	POWER	POWER	POWER
count	4337.000000	366.000000	293.000000
mean	11730.214760	9875.638474	12336.121780
std	3991.630810	6189.736377	4175.437200
min	2720.133922	0.000000	3302.594598
25%	8185.832753	5402.198359	7630.350973
50%	13473.368134	12824.136655	14580.370812
75%	14972.032226	15543.685624	15651.285589
max	17616.227967	16763.060437	16763.060437

How is the train Power vs test Power

What can we improve ?

- Make better implementations
- Look for different algorithms
- and ..