Baker Hughes Hackathon

Predictive Modeling (Team 2)

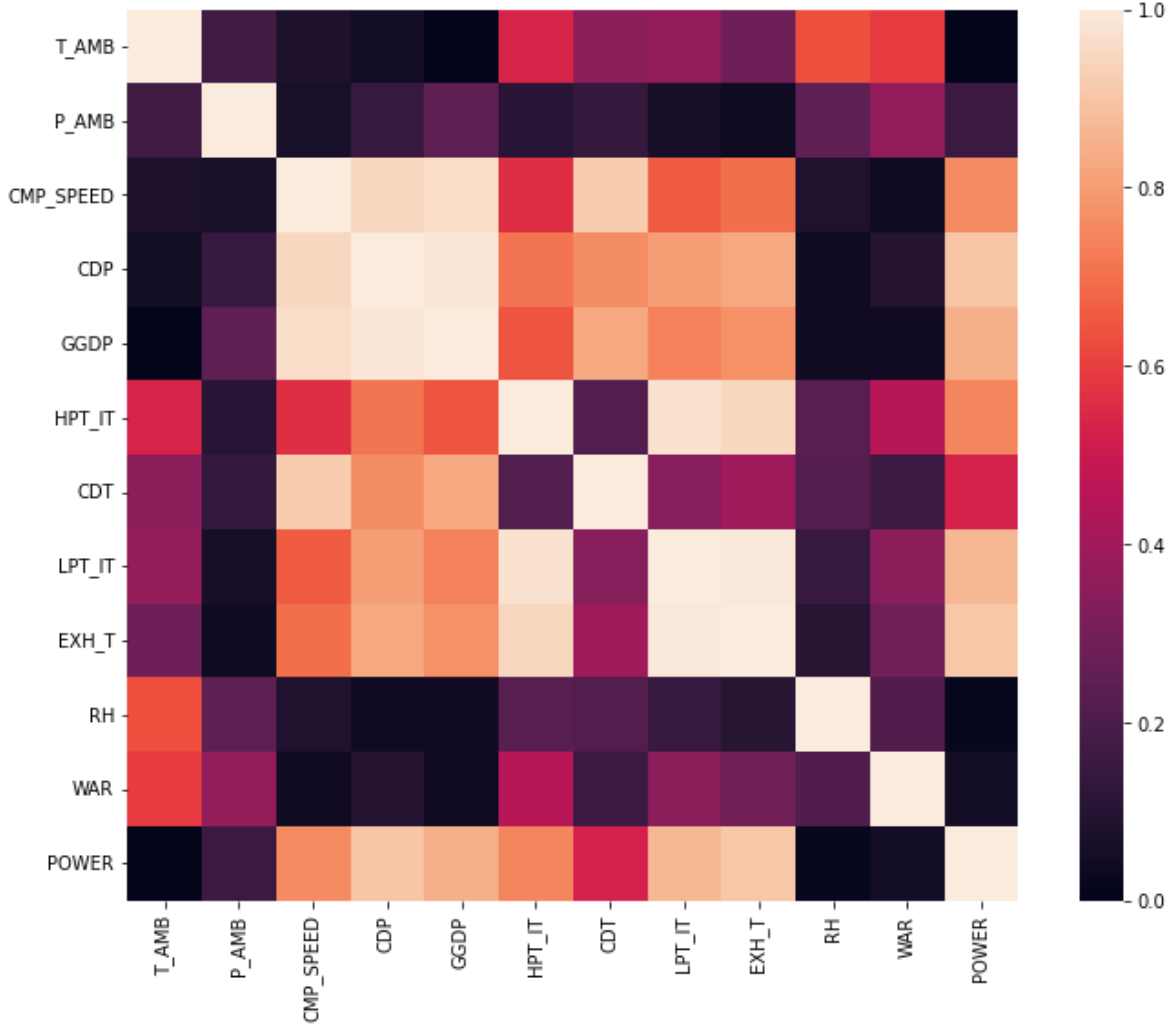
Isaura Ramírez Salazar
José David Romo López
Juan Diego Sanchez Díaz

Data Visualization

(all training data)

abs(CorrelationCoeff)

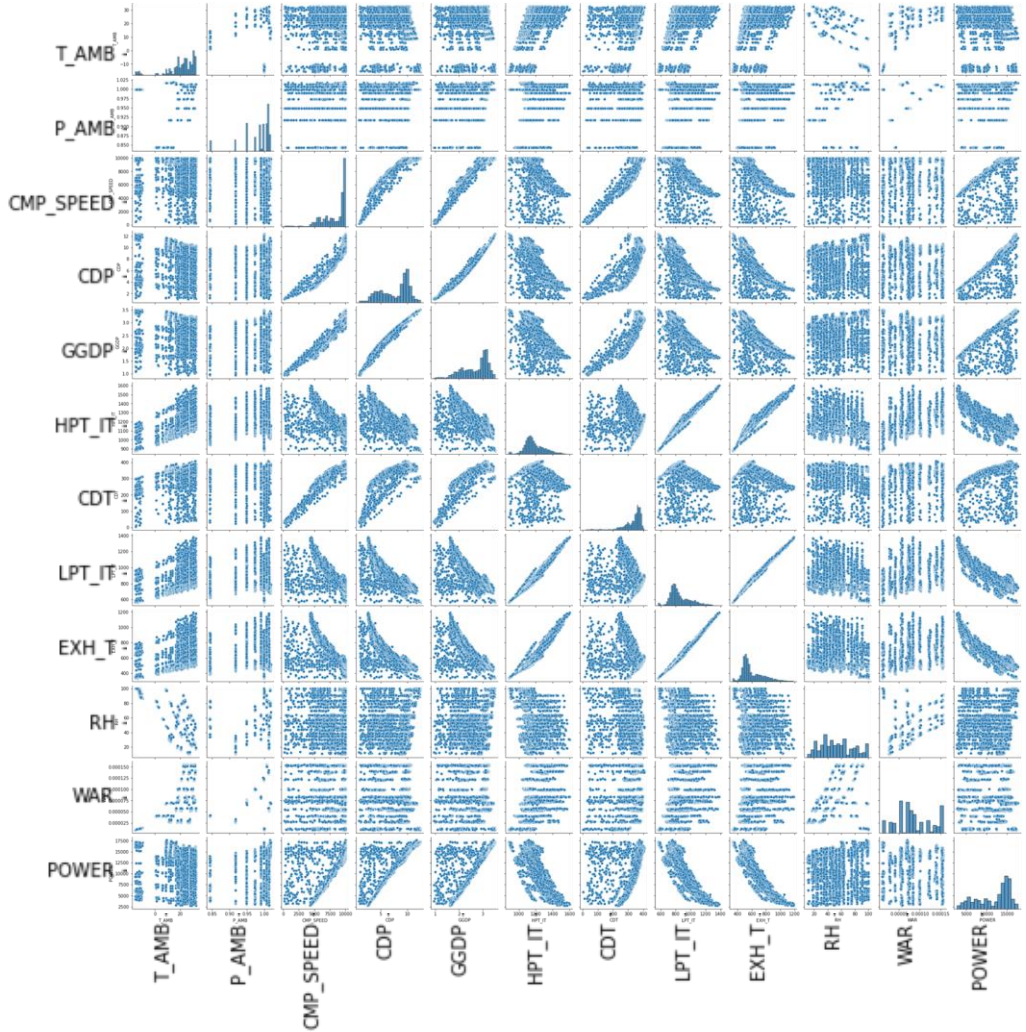
The clearer the better



Data Visualization

(all training data)

abs(CorrelationCoeff)
The clearer the better



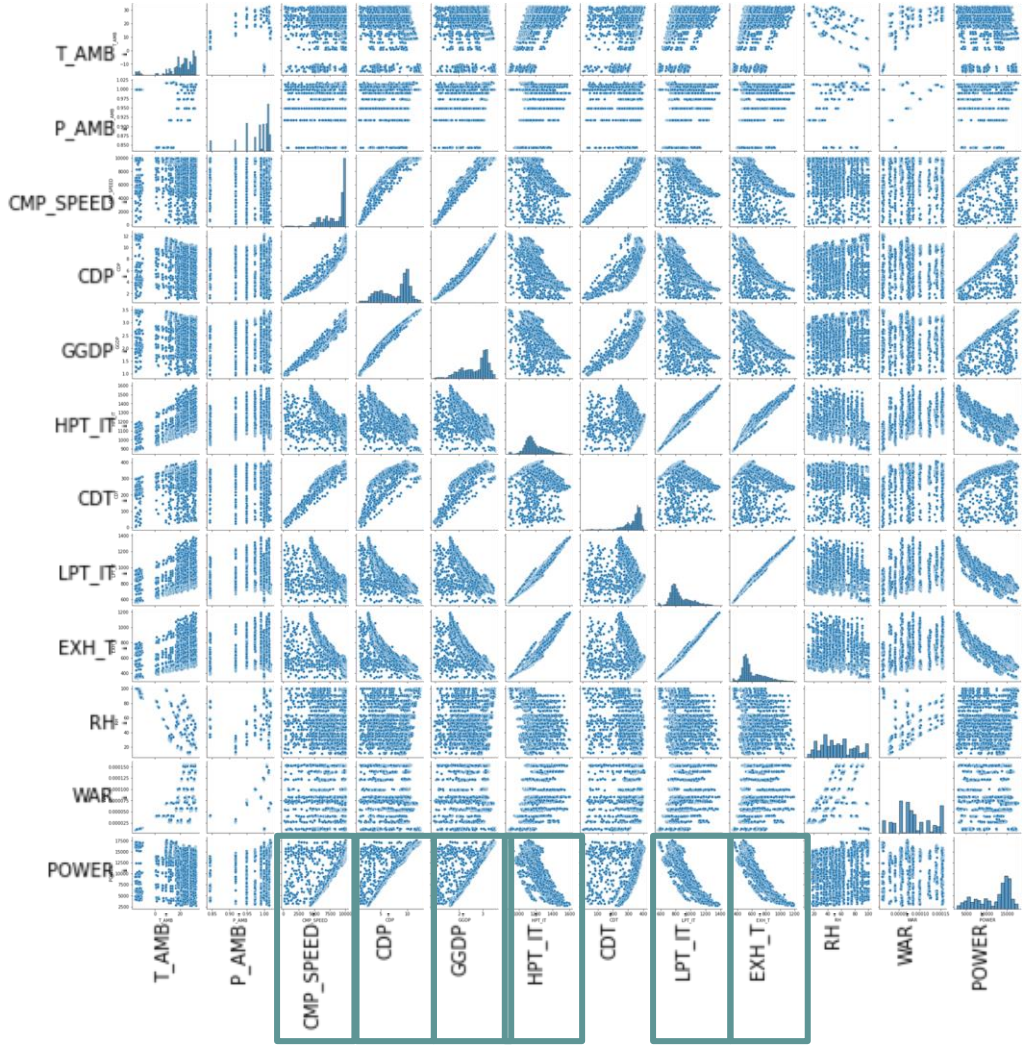
For us the interesting variables are in order:

```
POWER        1.000000
EXH_T        0.903272
CDP          0.898743
LPT_IT       0.863357
GGDP         0.846911
CMP_SPEED    0.758151
HPT_IT       0.746706
CDT          0.525526
P_AMB        0.155483
WAR          0.049835
RH           0.017927
T_AMB        0.004338
Name: POWER, dtype: float64
```

Data Visualization

(all training data)

abs(CorrelationCoeff)
The clearer the better

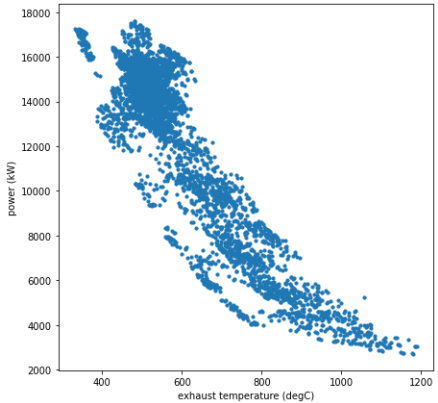


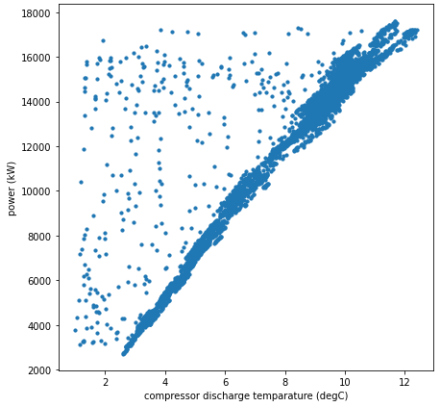
For us the interesting variables are in order:

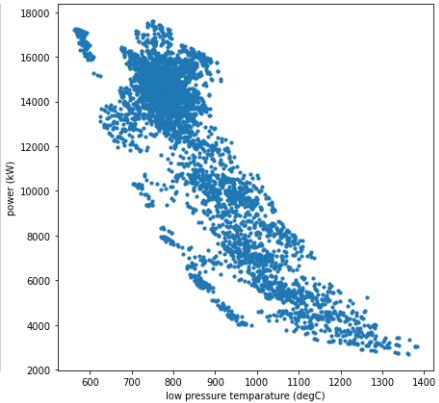
```
POWER          1.000000
EXH_T          0.903272
CDP            0.898743
LPT_IT         0.863357
GGDP           0.846911
CMP_SPEED      0.758151
HPT_IT         0.746706
CDT            0.525526
P_AMB          0.155483
WAR            0.049835
RH             0.017927
T_AMB          0.004338
Name: POWER, dtype: float64
```

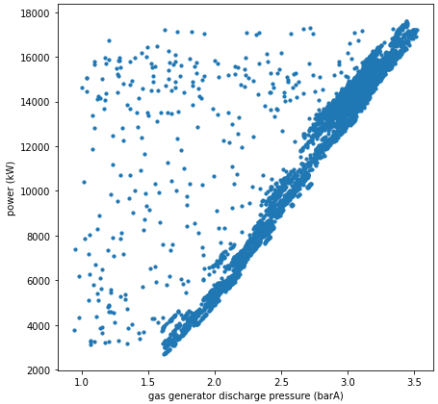
Data Visualization

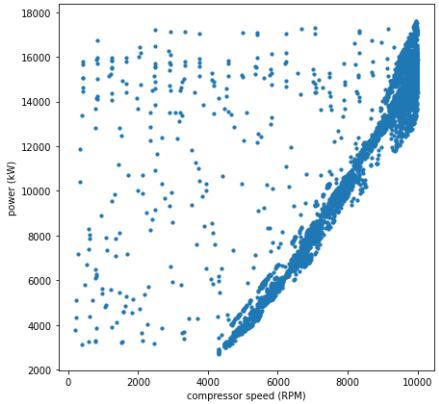
We visualize all the training data

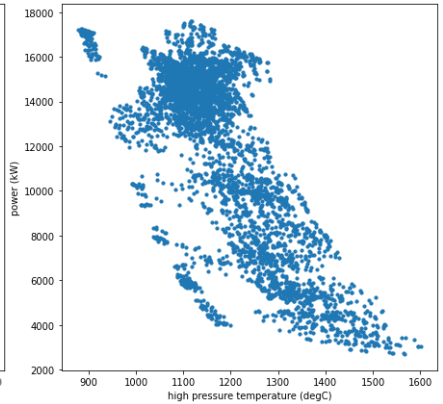













```
POWER          1.000000
EXH_T          0.903272
CDP            0.898743
LPT_IT         0.863357
GGDP           0.846911
CMP_SPEED      0.758151
HPT_IT         0.746706
CDT            0.525526
P_AMB          0.155483
WAR            0.049835
RH             0.017927
T_AMB          0.004338
Name: POWER, dtype: float64
```

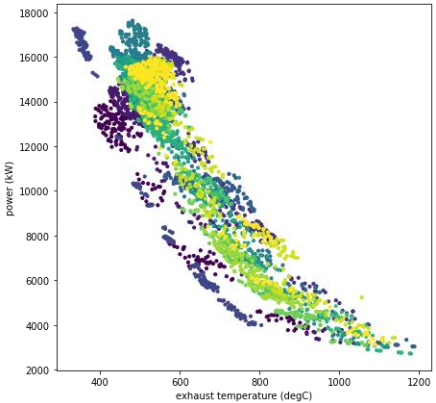
Data Visualization

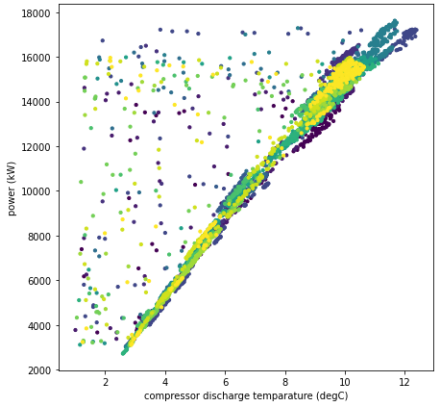
Maybe if we separate distint files. Could be a metadata depandence?

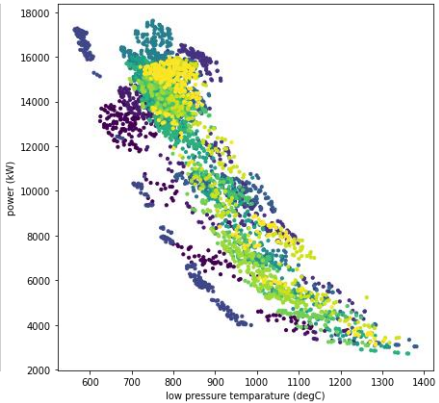
Doesn't look very important.

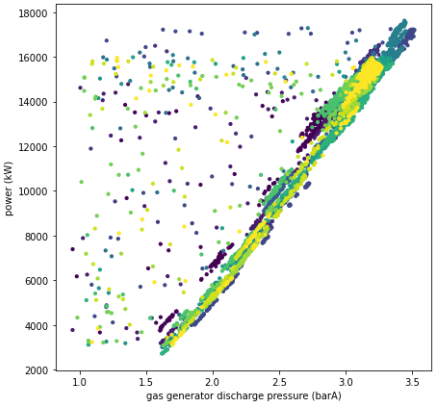
POWER	1.000000
EXH_T	0.903272
CDP	0.898743
LPT_IT	0.863357
GGDP	0.846911
CMP_SPEED	0.758151
HPT_IT	0.746706
CDT	0.525526
P_AMB	0.155483
WAR	0.049835
RH	0.017927
T_AMB	0.004338

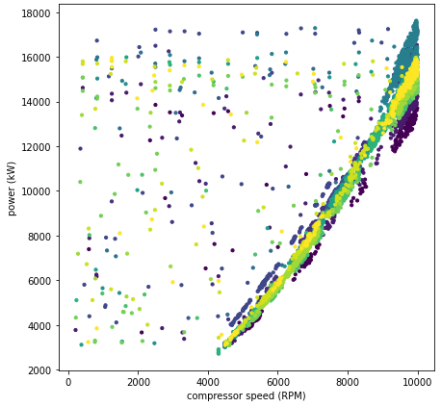
Name: POWER, dtype: float64

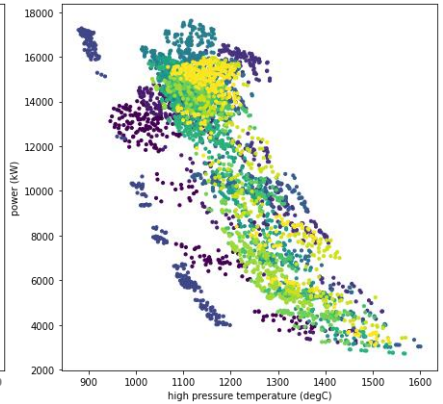










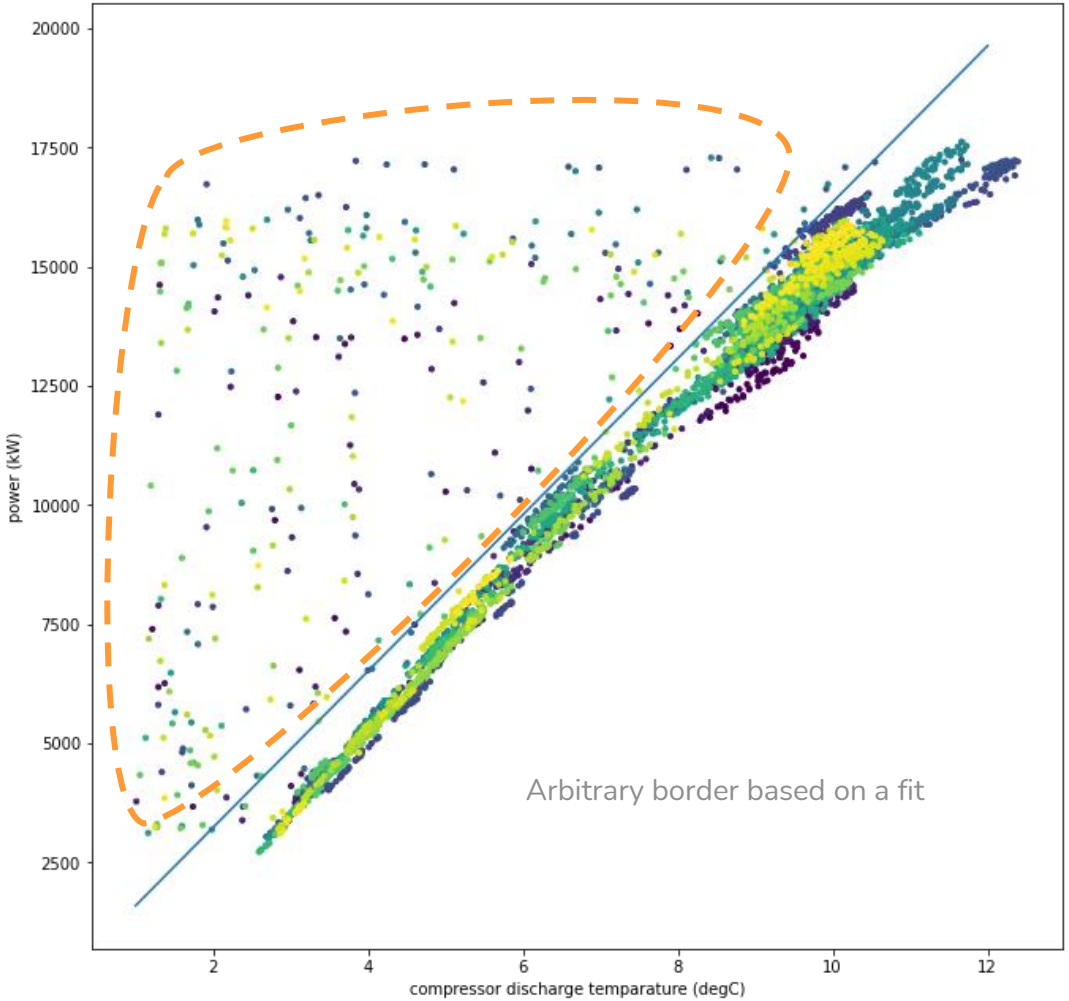


Data Visualization

‘Data’ + **‘Outliers’**

Data: Below the rect
‘Outliers’: Above the rect

The dependence power vs CDT looks like a linear dependence except from the presence of some ’outliers’

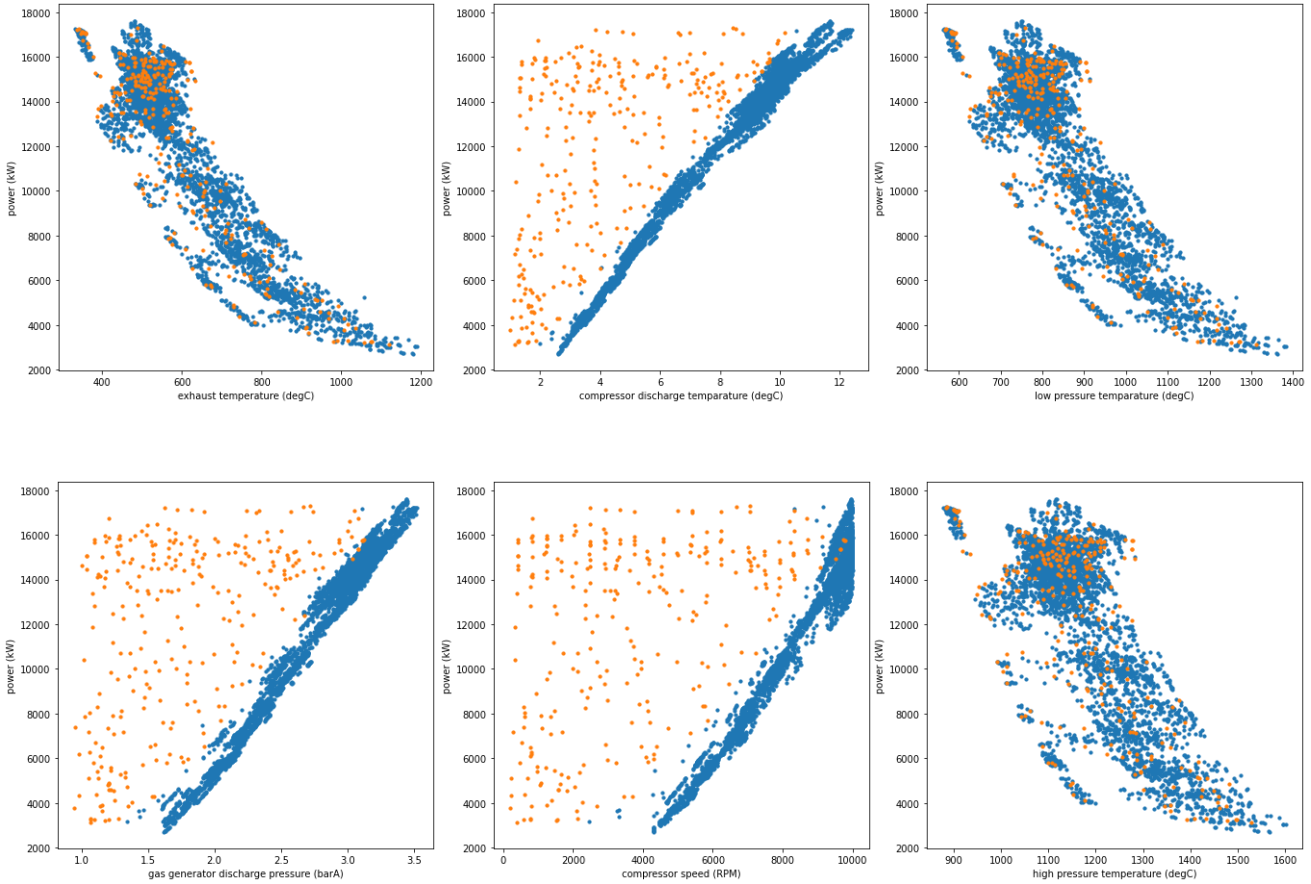


Data Visualization

'Data' + 'Outliers'

But, are they really outliers?

- ❑ They show a consistent behaviour.
- ❑ They are not <u>rare</u> ~6% of the sample

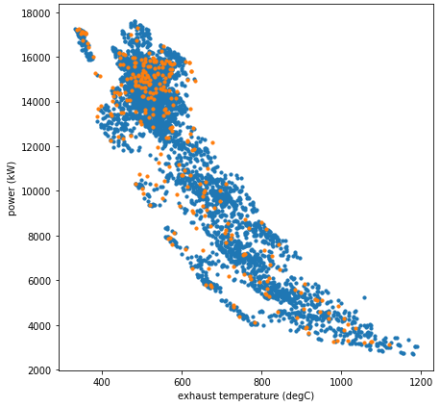


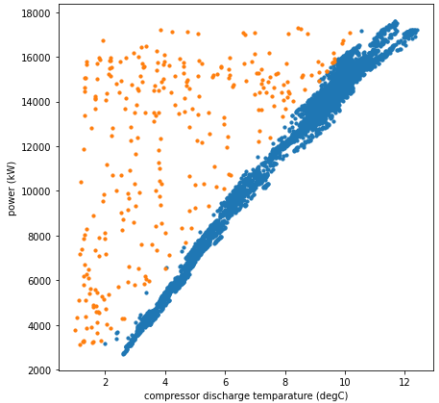
Data Visualization

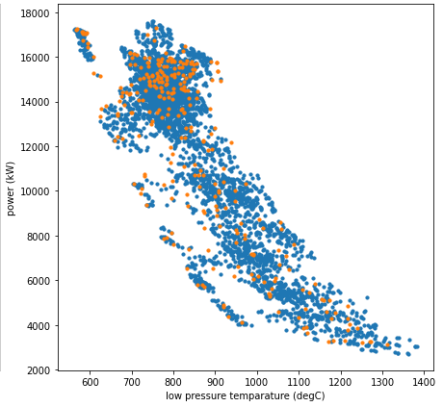
~~'Data'~~ + ~~'Outliers'~~

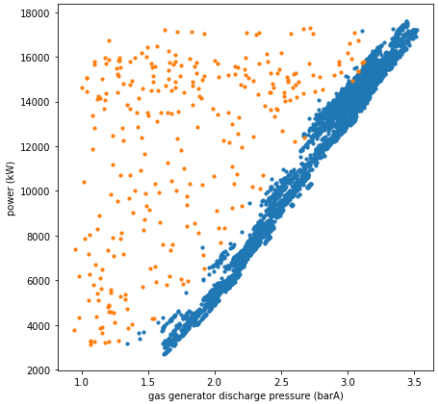
Blue type + **Orange type**

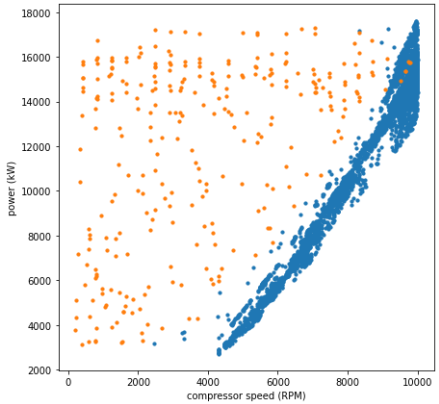
Each with a different estimator method

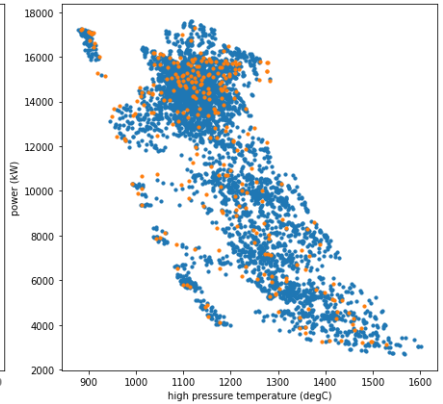








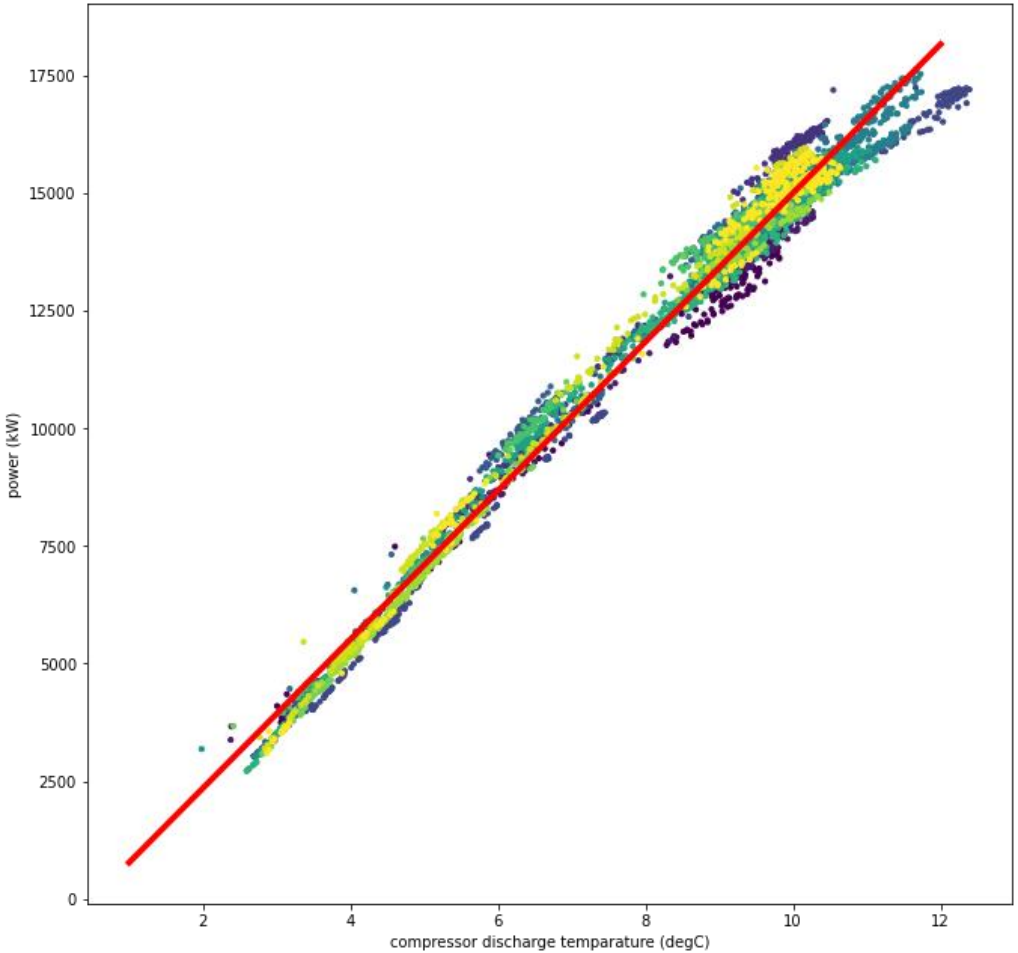




Predictors

Blue predictor

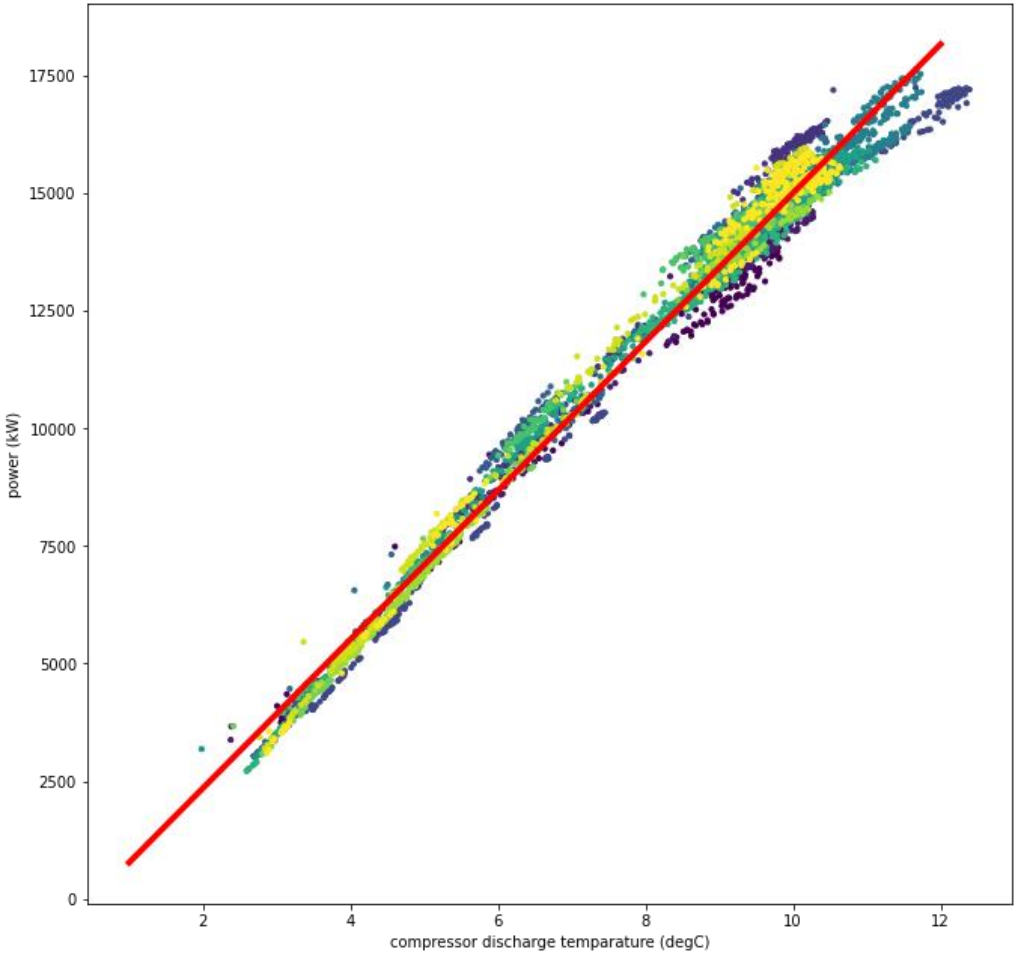
Linear Regression



Predictors

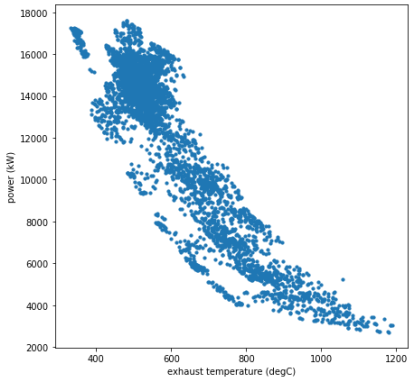
Orange predictor

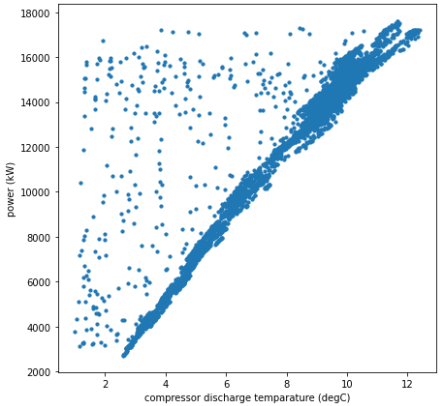
Linear Regression

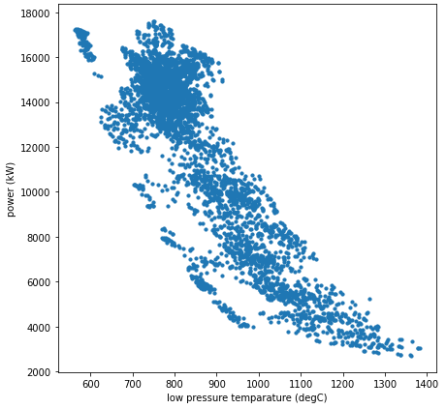


Data Visualization

Data Visualization



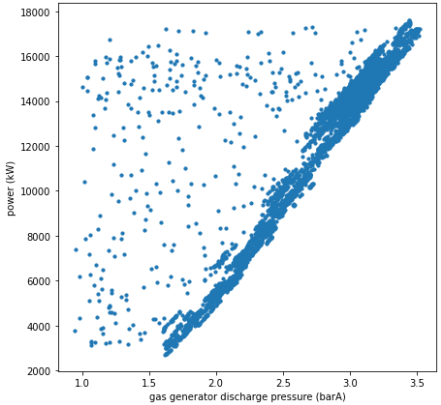


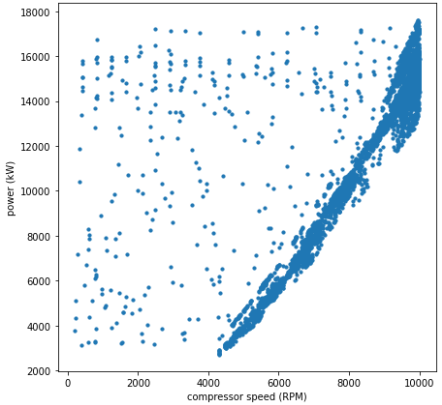


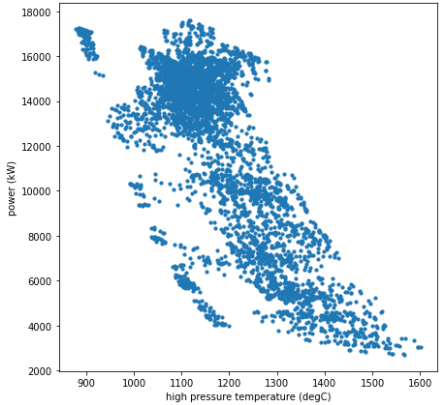
For us the interesting variables are in order:

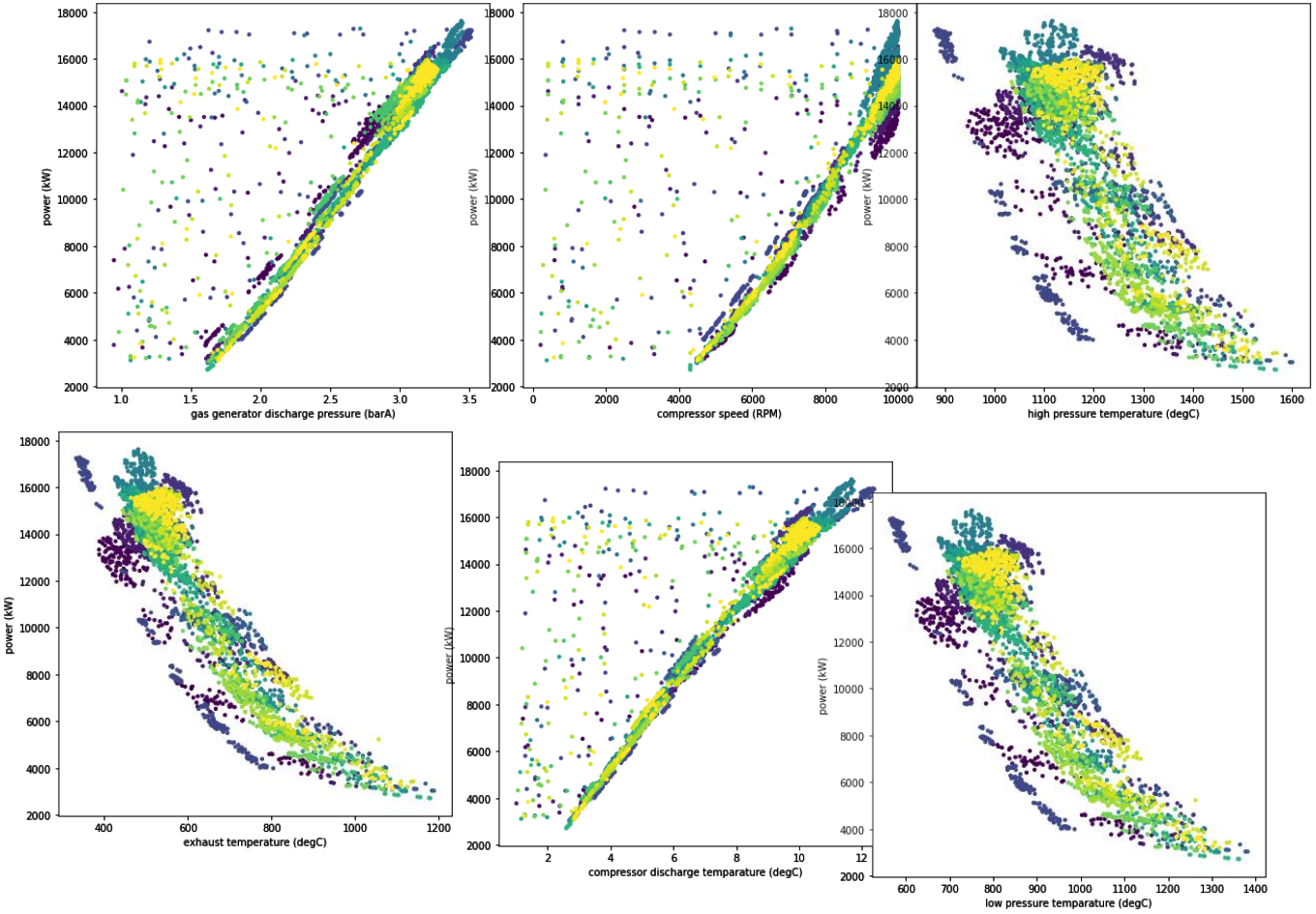
POWER	1.000000
EXH_T	0.903272
CDP	0.898743
LPT_IT	0.863357
GGDP	0.846911
CMP_SPEED	0.758151
HPT_IT	0.746706
CDT	0.525526
P_AMB	0.155483
WAR	0.049835
RH	0.017927
T_AMB	0.004338

Name: POWER, dtype: float64








power (kW)
gas generator discharge pressure (barA)
compressor speed (RPM)
high pressure temperature (degC)
exhaust temperature (degC)
compressor discharge temparature (degC)
low pressure temparature (degC)