Code to clean

The following data was used to make a perfect dataset out of the raw dataset by cleaning, merging, joining and dropping unnecessary columns.

In [57]:

```
import pandas as pd
import numpy as np
```

In [58]:

```
plant1_gen=pd.read_csv('Dataset/P1G.csv')
plant1_weat=pd.read_csv('Dataset/P1W.csv')
```

In [59]:

```
plant1_gen = plant1_gen.groupby('DATE_TIME').agg({'DC_POWER':'mean','AC_POWER':'mean', '
DAILY_YIELD':'mean','TOTAL_YIELD':'mean'})
plant1_weat=plant1_weat.set_index('DATE_TIME', drop=True)
```

In [60]:

```
plant1_gen
```

Out[60]:

DATE_TIME	DC_POWER	AC_POWER	DAILY_YIELD	TOTAL_YIELD
01-06-2020 00:00	0.0	0.0	245.784091	6.978158e+06
01-06-2020 00:15	0.0	0.0	0.000000	6.978158e+06
01-06-2020 00:30	0.0	0.0	0.000000	6.978158e+06
01-06-2020 00:45	0.0	0.0	0.000000	6.978158e+06
01-06-2020 01:00	0.0	0.0	0.000000	6.978158e+06
...	...	...	...	...
31-05-2020 22:45	0.0	0.0	5695.045455	6.978158e+06
31-05-2020 23:00	0.0	0.0	5695.045455	6.978158e+06
31-05-2020 23:15	0.0	0.0	5695.045455	6.978158e+06
31-05-2020 23:30	0.0	0.0	5695.045455	6.978158e+06
31-05-2020 23:45	0.0	0.0	5169.870130	6.978158e+06

3158 rows × 4 columns

In [61]:

```
plant1_weat
```

Out[61]:

DATE_TIME	PLANT_ID	SOURCE_KEY	AMBIENT_TEMPERATURE	MODULE_TEMPERATURE	IRRADIATION
15-05-2020 00:00	4135001	HmiyD2TTLFNqkNe	25.184316	22.857507	0.0
15-05-2020 00:15	4135001	HmiyD2TTLFNqkNe	25.084589	22.761668	0.0
15-05-2020 00:30	4135001	HmiyD2TTLFNqkNe	24.935753	22.592306	0.0
15-05-2020 00:45	4135001	HmiyD2TTLFNqkNe	24.846130	22.360852	0.0
15-05-2020 01:00	4135001	HmiyD2TTLFNqkNe	24.621525	22.165423	0.0
...	...	...	...	...	...
17-06-2020 22:45	4135001	HmiyD2TTLFNqkNe	22.150570	21.480377	0.0
17-06-2020 23:00	4135001	HmiyD2TTLFNqkNe	22.129816	21.389024	0.0
17-06-2020 23:15	4135001	HmiyD2TTLFNqkNe	22.008275	20.709211	0.0
17-06-2020 23:30	4135001	HmiyD2TTLFNqkNe	21.969495	20.734963	0.0
17-06-2020 23:45	4135001	HmiyD2TTLFNqkNe	21.909288	20.427972	0.0

3182 rows × 5 columns

In [62]:

```
plant1=pd.merge(plant1_gen, plant1_weat, how='inner', left_index=True, right_index=True)
df=plant1
df=df.reset_index(drop=False, inplace=False)
```

In [63]:

```
df=df.drop(labels=['SOURCE_KEY','PLANT_ID','TOTAL_YIELD', 'DAILY_YIELD', 'AC_POWER','DAT
E_TIME'], axis=1)
```

In [64]:

```
df.to_csv('Dataset/P1.csv')
df
```

Out[64]:

	DC_POWER	AMBIENT_TEMPERATURE	MODULE_TEMPERATURE	IRRADIATION
0	0.0	23.128673	20.464305	0.0
1	0.0	23.032562	20.341429	0.0
2	0.0	22.967493	20.269493	0.0
3	0.0	22.810594	20.198918	0.0
4	0.0	22.611436	20.085866	0.0
...	...	...	...	...
3152	0.0	23.670292	21.691071	0.0
3153	0.0	23.795434	22.067778	0.0
3154	0.0	23.727901	21.662972	0.0
3155	0.0	23.497284	21.051402	0.0
3156	0.0	23.244698	20.774560	0.0

3157 rows × 4 columns

In [65]:

```
df
```

Out[65]:

	DC_POWER	AMBIENT_TEMPERATURE	MODULE_TEMPERATURE	IRRADIATION
0	0.0	23.128673	20.464305	0.0
1	0.0	23.032562	20.341429	0.0
2	0.0	22.967493	20.269493	0.0
3	0.0	22.810594	20.198918	0.0
4	0.0	22.611436	20.085866	0.0
...	...	...	...	...
3152	0.0	23.670292	21.691071	0.0
3153	0.0	23.795434	22.067778	0.0
3154	0.0	23.727901	21.662972	0.0
3155	0.0	23.497284	21.051402	0.0
3156	0.0	23.244698	20.774560	0.0

3157 rows × 4 columns

In [ ]:

```

```