3. EMC Data Science Global Hackathon (Air Quality Prediction)

Build a local early warning systems to accurately predict dangerous levels of air pollutants on an hourly basis.

https://www.kaggle.com/c/dsg-hackathon#description

The challenge for the hackathon is to build with better more accurate predictive models of metropolitan air pollution. The EPA's Air Quality Index is used daily by people suffering from asthma and other respiratory diseases to avoid dangerous levels of outdoor air pollutants, which can trigger attacks. According to the World Health Organisation there are now estimated to be 235 million people suffering from asthma. Globally, it is now the most common chronic disease among children, with incidence in the US doubling since 1980. The model we build could be used as the basis for an early warning system that is capable of accurately predicting dangerous levels of air pollutants on an hourly basis.

Data Science Global is a non-profit organization dedicated to bringing together the world's communities of data scientists, artists, technologists and visionaries. For our inaugural event, we are hosting a global data science hackathon. It will be taking place simultaneously in cities around the world: London, New York, Boston, Chicago, San Francisco, Melbourne, Canberra, Sydney and Turku, Finland, as well as remote participants competing directly through Kaggle. You can join in the live webcast from the participating venues at datascienceglobal.org

Hourly data on various targets were collected for 8 days from various sites around Chicago was provided as training data. The aim of the challenge is to predict various time points within the next 3 days after the training period (1, 2 ,3, 4, 5, 10, 17, 24, 48, and 72 hours after the end of the 8-day training data).

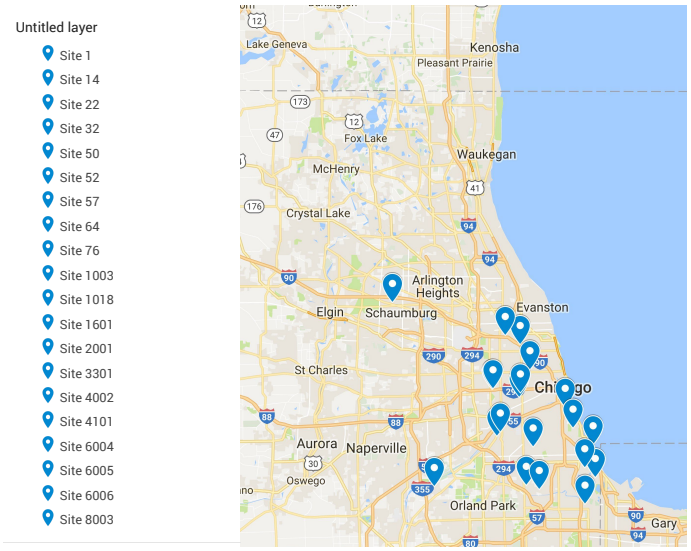
Summary of the data on Targets and Sites:

		Sites																		
		1	14	22	32	50	52	57	64	76	1003	1018	1601	2001	3301	4002	4101	6004	6005	6006	8003
Targets	target_1							x													
	target_2							x													
	target_3	x				x		x					x			x				x	
	target_4	x				x		x				x	x	x		x	x			x	x
	target_5																			x	
	target_7							x													
	target_8							x								x		x			x
	target_9															x					x
	target_10															x					x
	target_11	x			x	x			x		x		x			x					x
	target_14															x					x
	target_15							x													

Target Code:

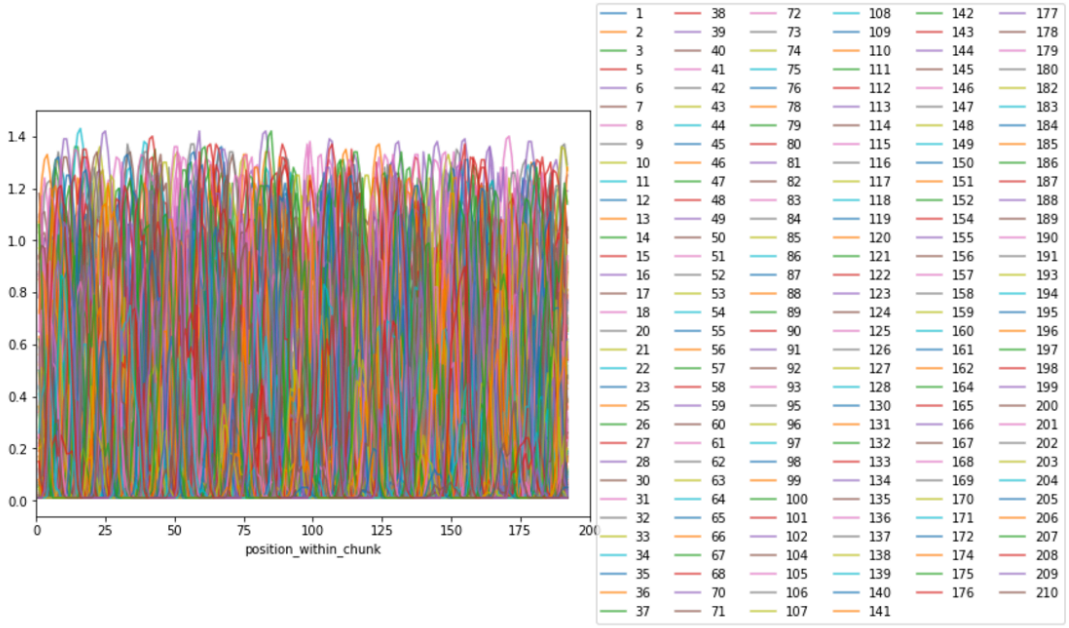
measured_quantity	PARAMETER_DESC
target_8	Carbon monoxide
target_4	Sulfur dioxide
target_3	SO2 max 5-min avg
target_10	Nitric oxide (NO)
target_14	Nitrogen dioxide (NO2)
target_9	Oxides of nitrogen (NOx)
target_11	Ozone
target_5	PM10 Total 0-10um STP
target_15	OC CSN Unadjusted PM2.5
target_2	Total Nitrate PM2.5 LC
target_1	EC CSN PM2.5 LC TOT
target_7	Total Carbon PM2.5 LC TC
target_8	Sulfate PM2.5 LC

Mapping of the Sites in EMC Data Set.



There are 208 chunks of data, each chuck with 192 positions corresponding to 8 days of data. However, each chunk has a different start time, making it hard to see time of day or day of week pattern.

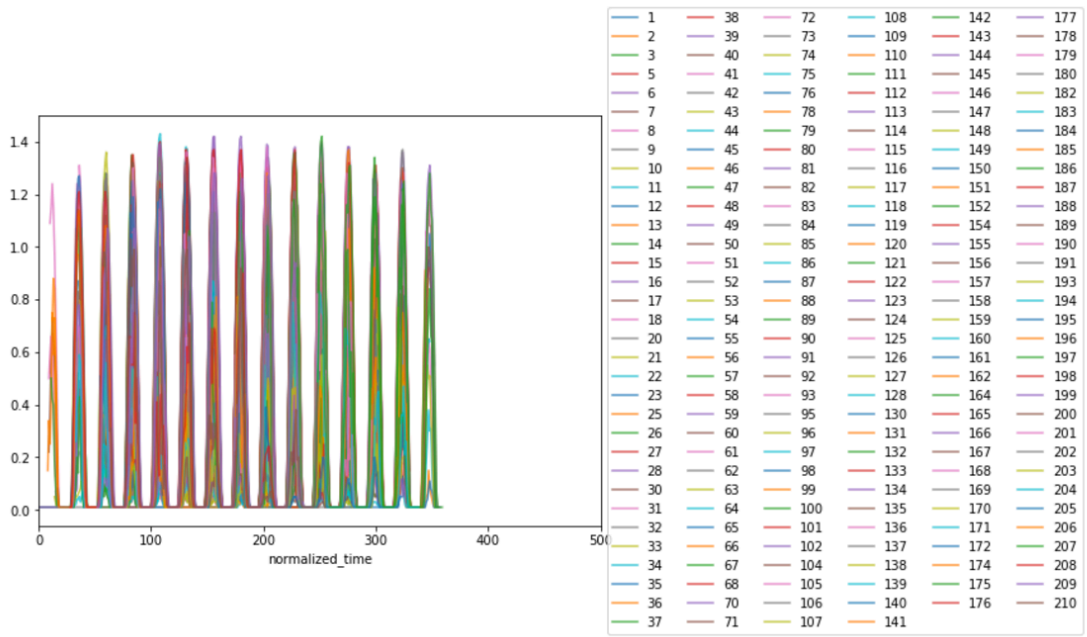
Solar Radiation



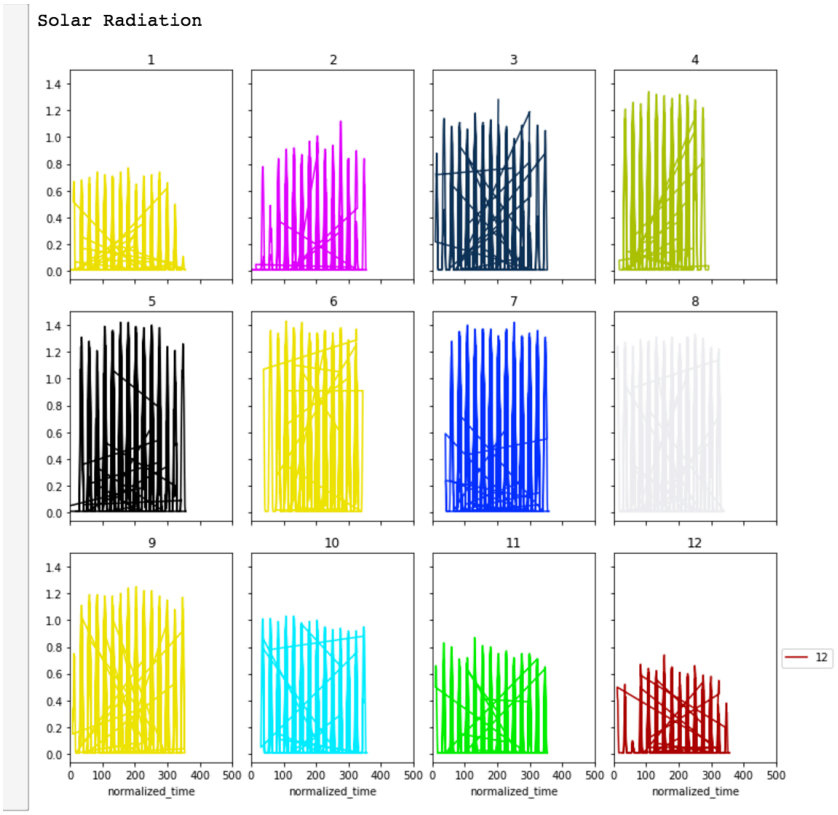
Need to normalized the time course.

Standardized	0	24	48	72	96	120	144	168	192	216	240	264	288	312	336	360	384
Hour	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00	0:00
Day	Sun	Mon	Tue	Wed	Thr	Fri	Sat	Sun	Sun	Mon	Tue	Wed	Thr	Fri	Sat	Sun	Mon

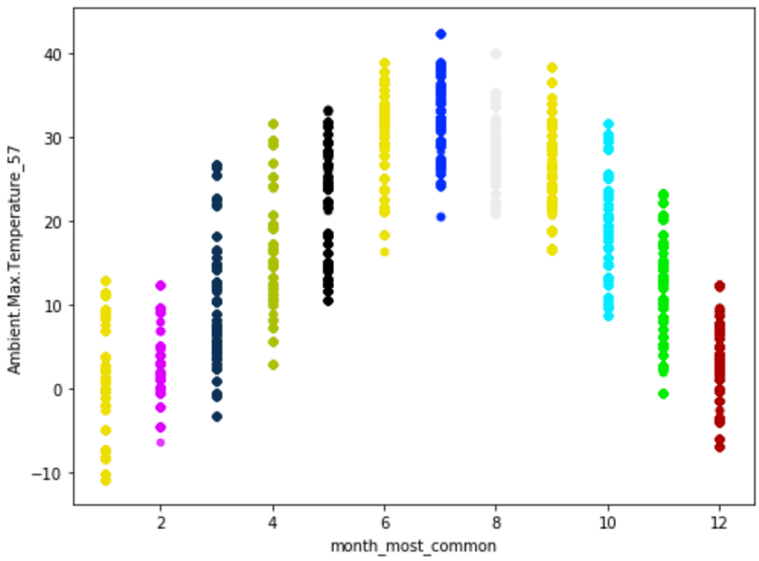
Radiation Data using new standardized time.



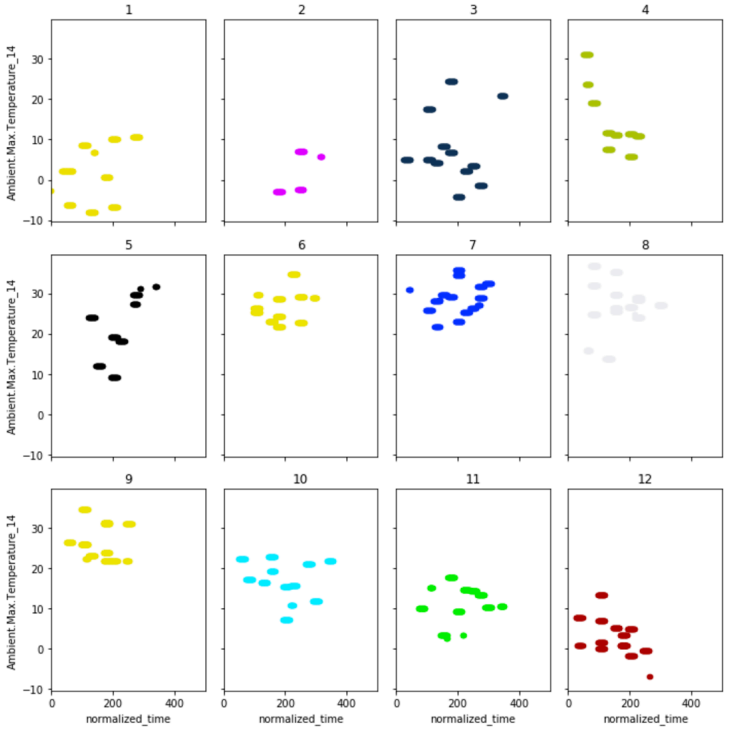
Breaking down the radiation data vs. standardized time by months:

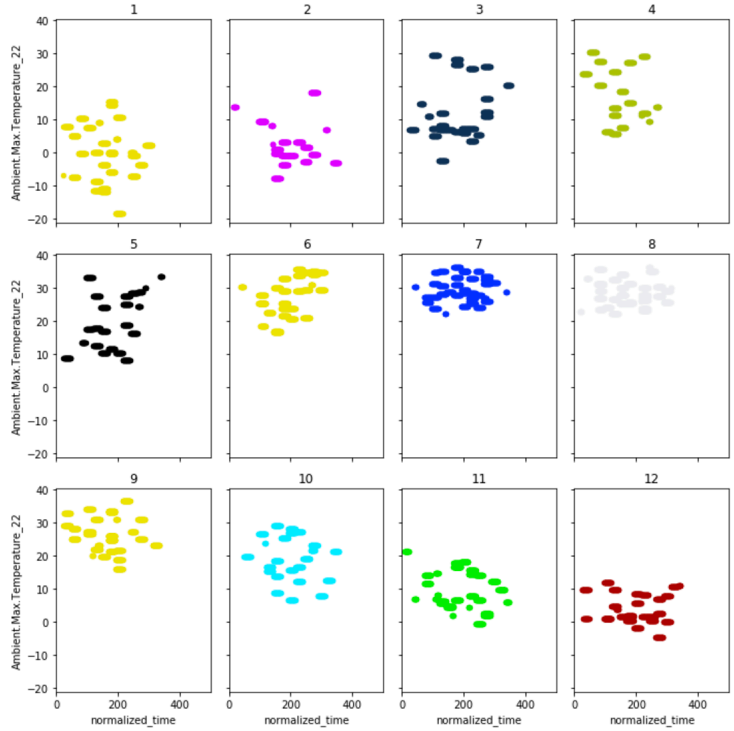


Max Temperature by Month:

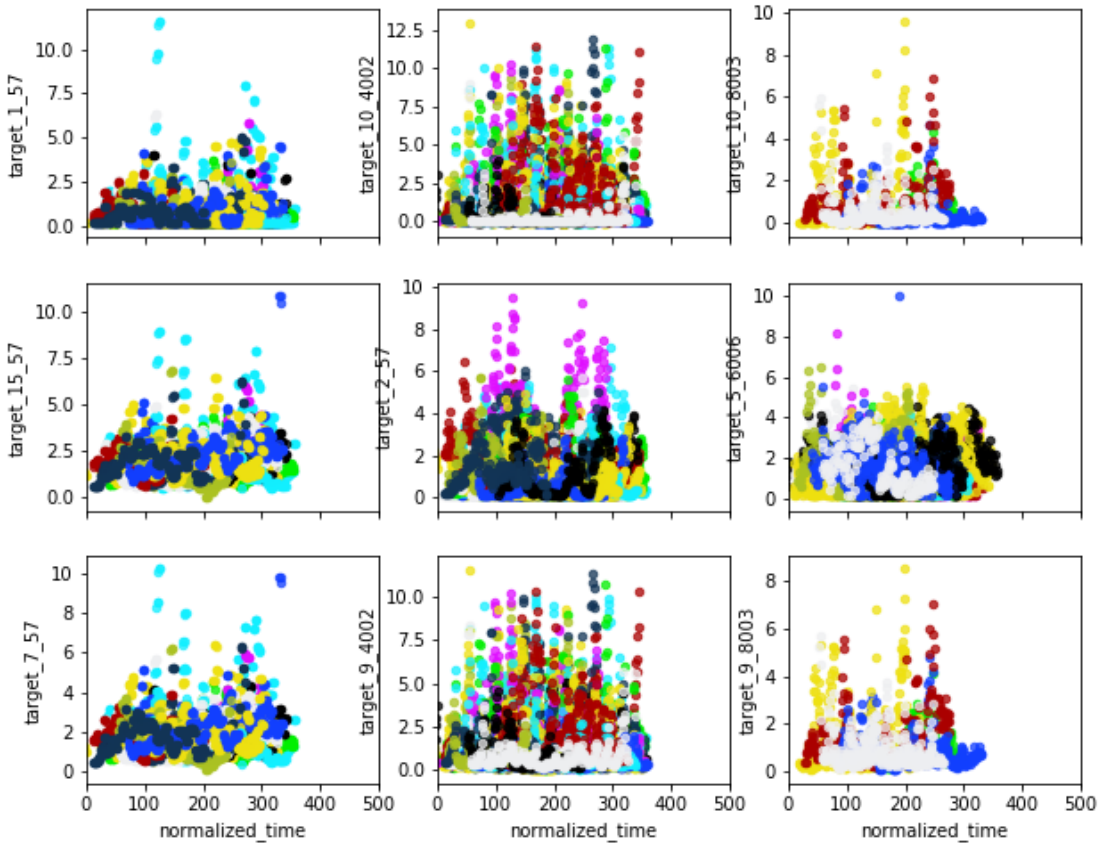


Max Temperature vs. Standardized time by Month in 2 Sites:





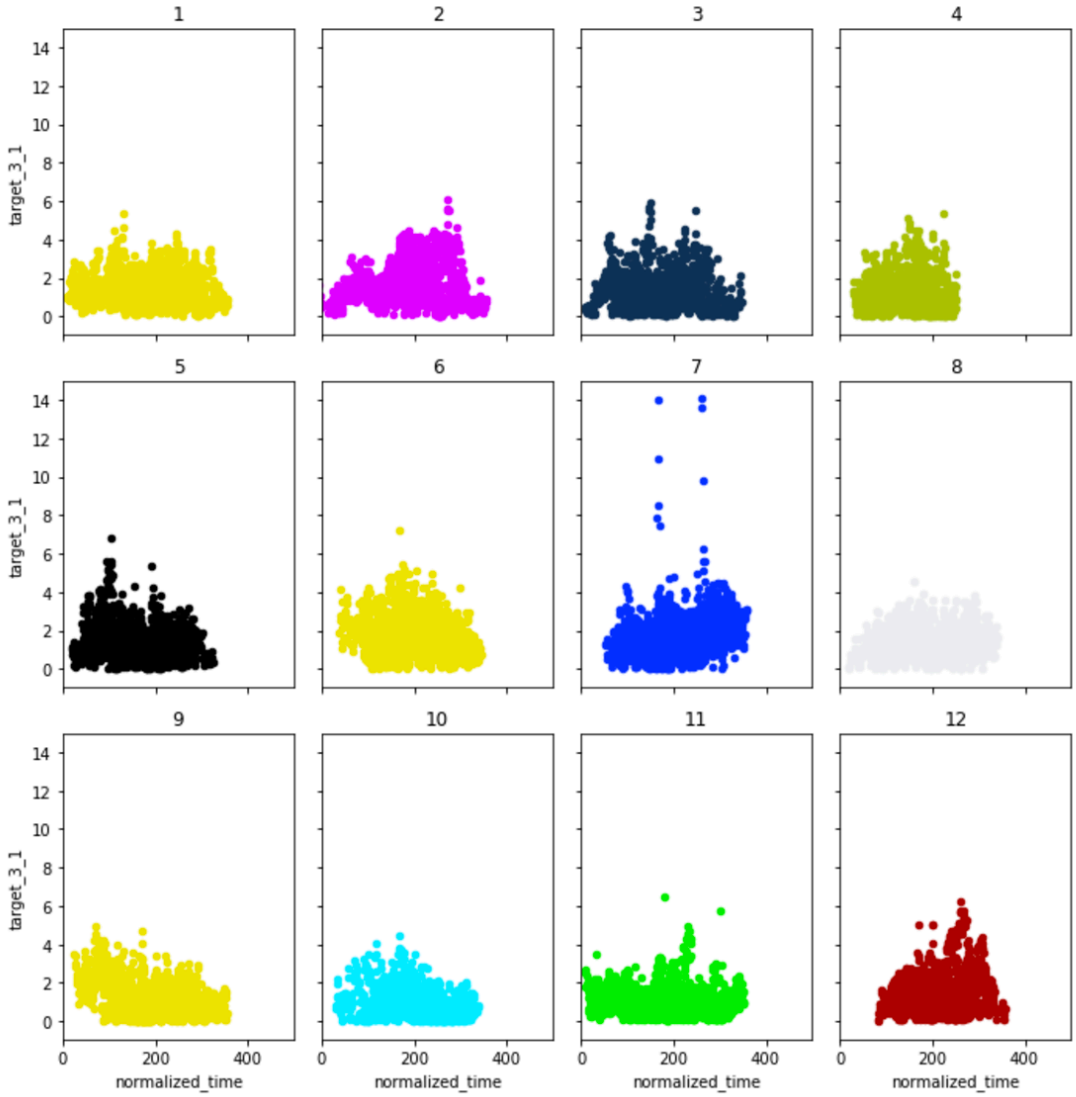
Plotting all chunks for each targets vs. standardized time. Color by chunks.



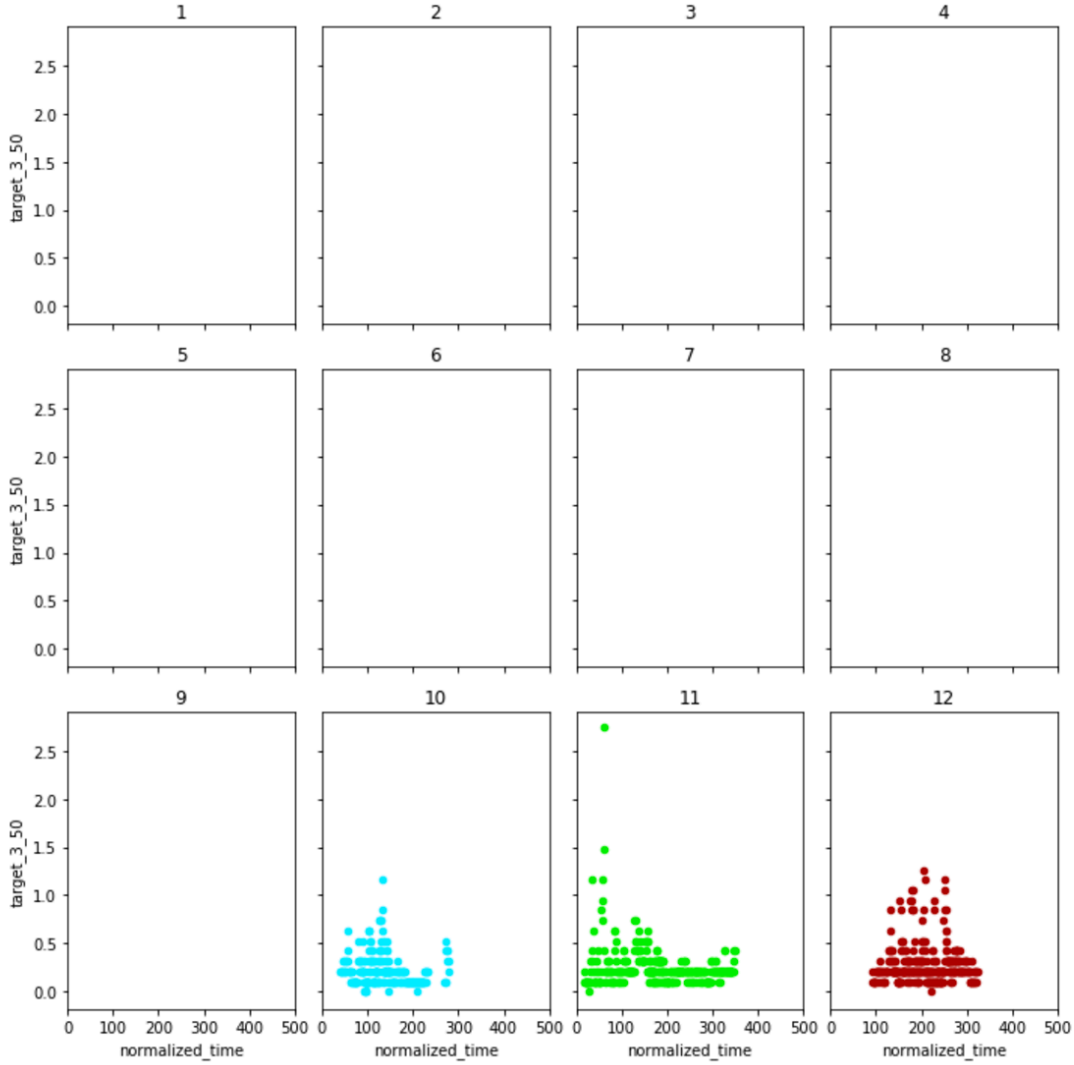
Conclusion: Too crowded. Need to separate chuncks by month.

Target 3:

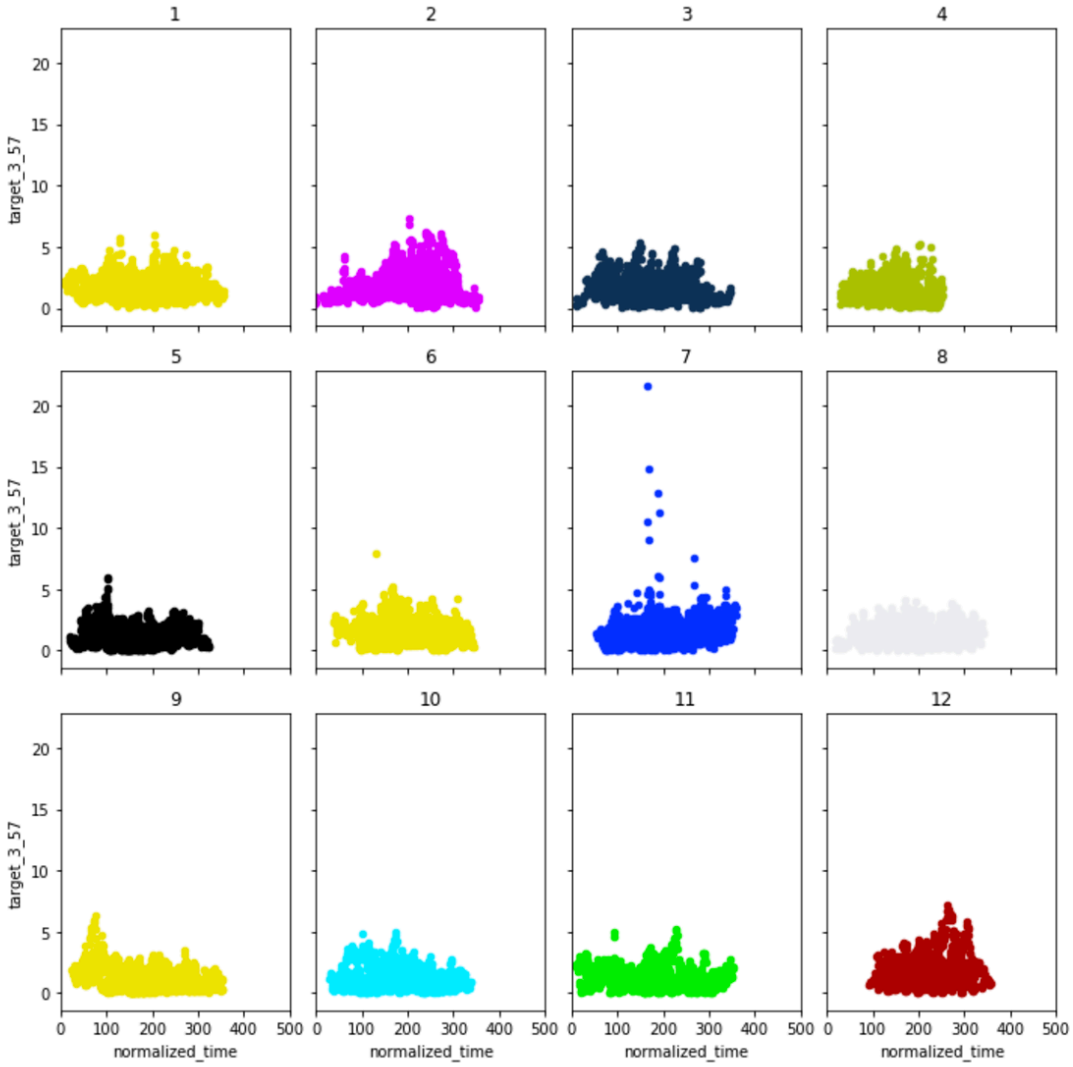
target_3_1



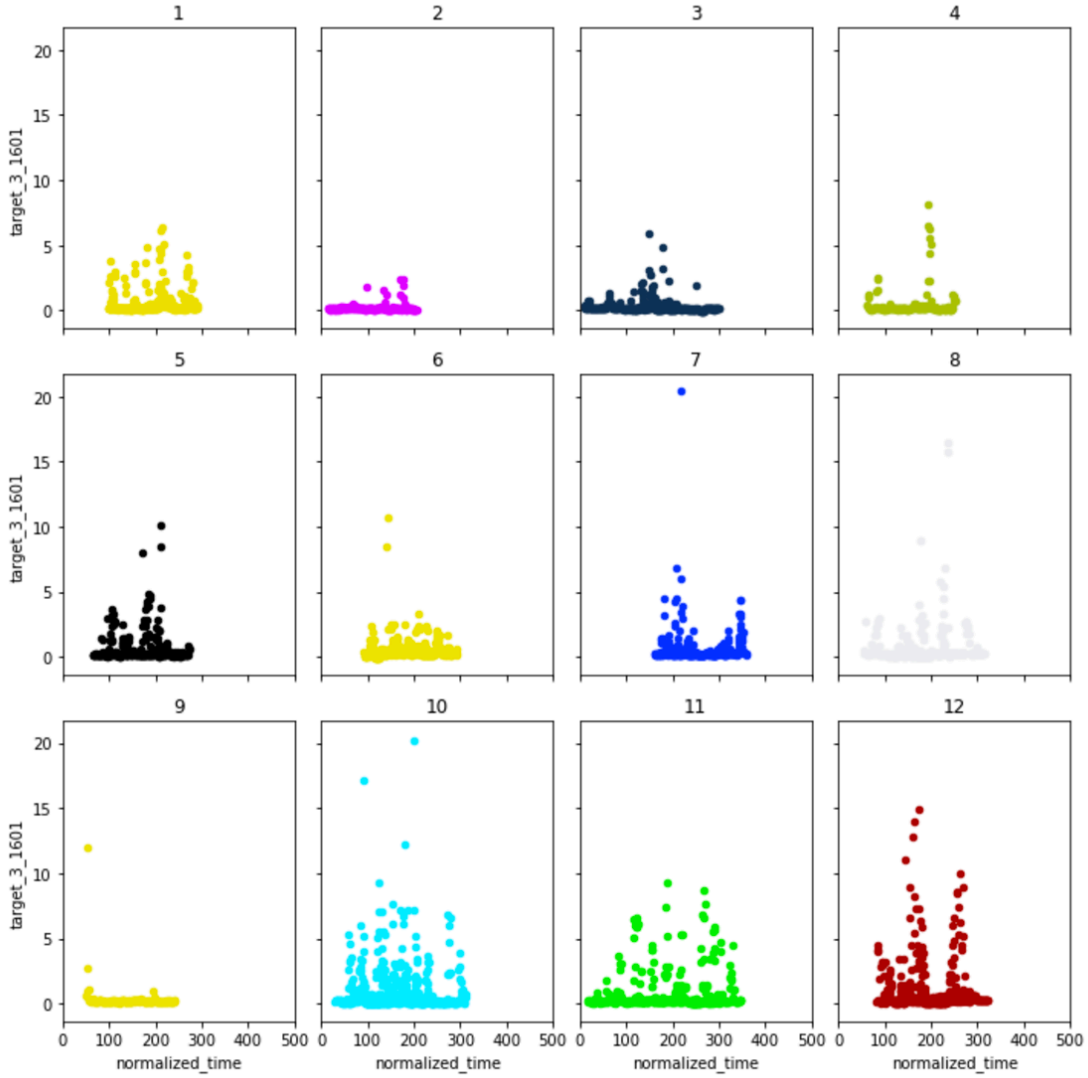
target_3_50


1
2
3
4
5
6
7
8
9
10
11
12
target_3_50
0.0
0.5
1.0
1.5
2.0
2.5
0
100
200
300
400
500
normalized_time

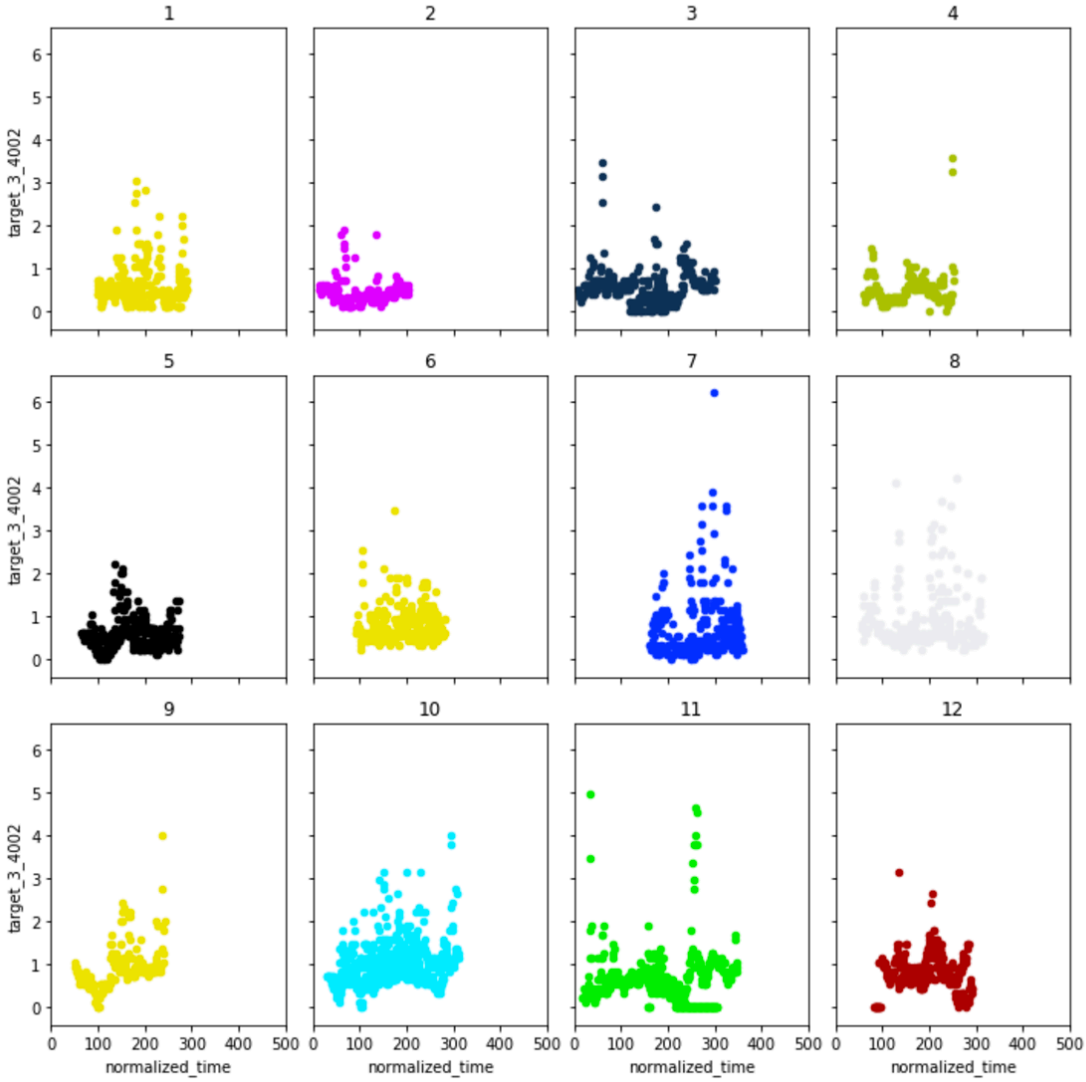
target_3_57


1
2
3
4
5
6
7
8
9
10
11
12
target_3_57
normalized_time
0
5
10
15
20
100
200
300
400
500

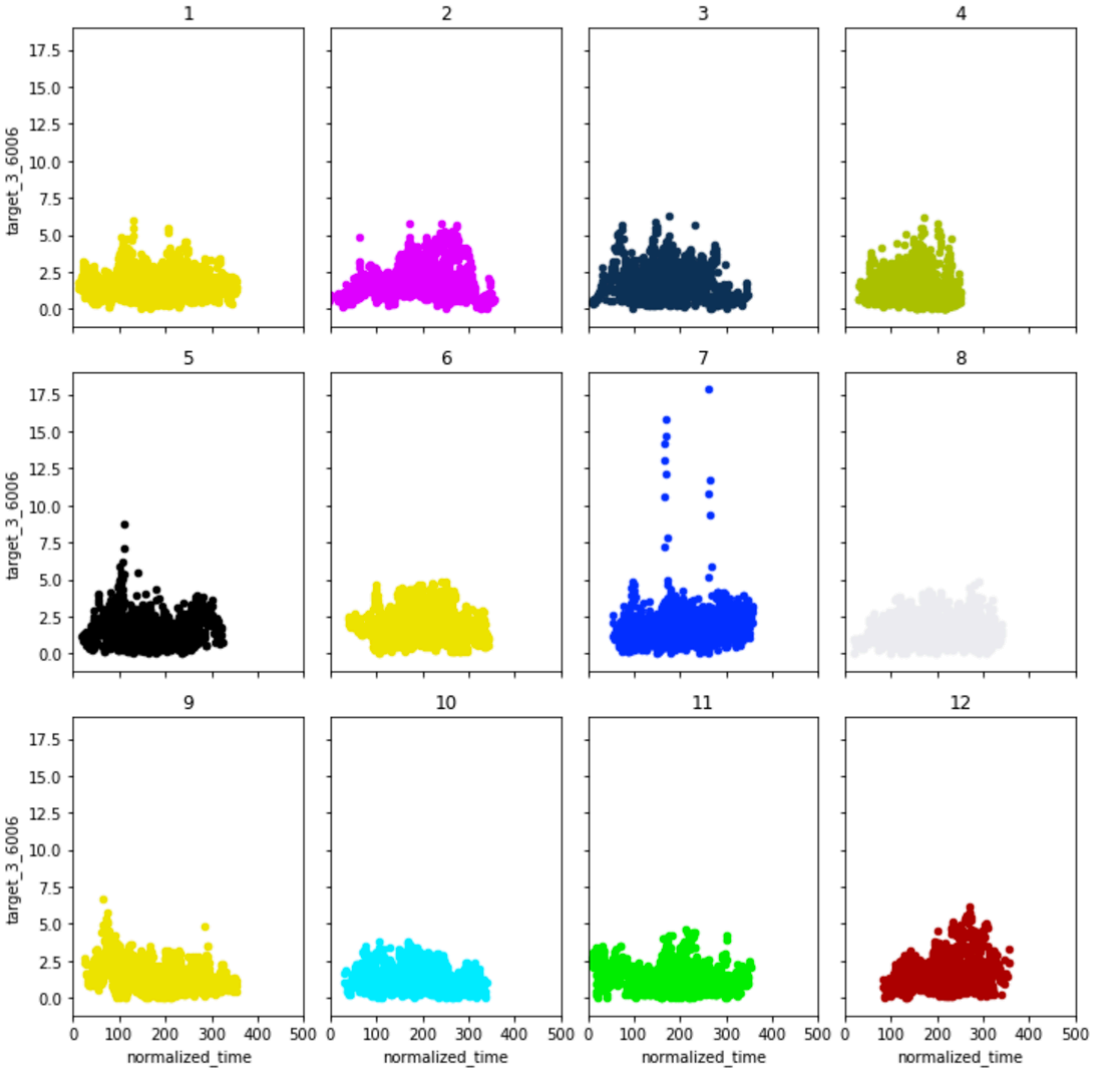
target_3_1601



target_3_4002

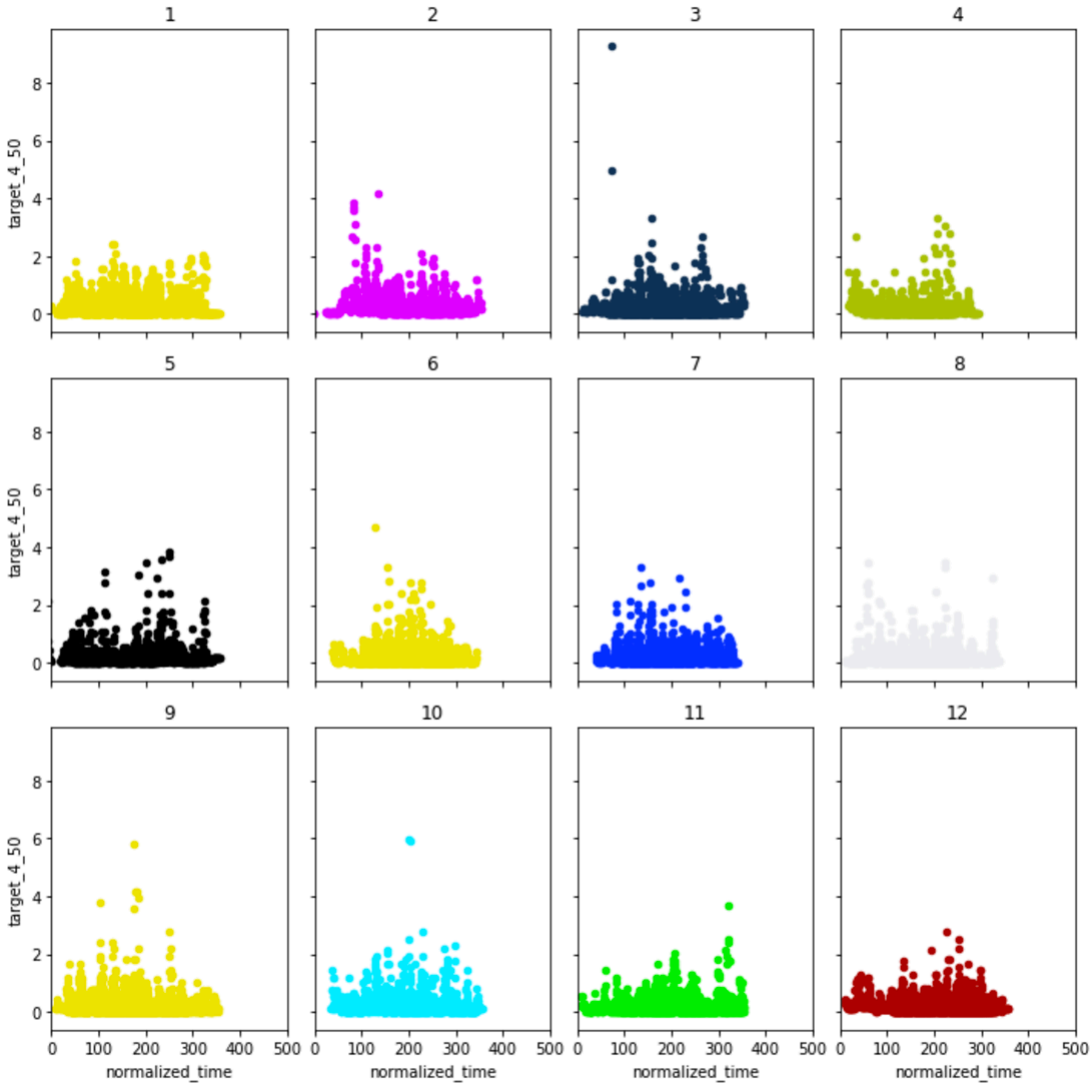


target_3_6006

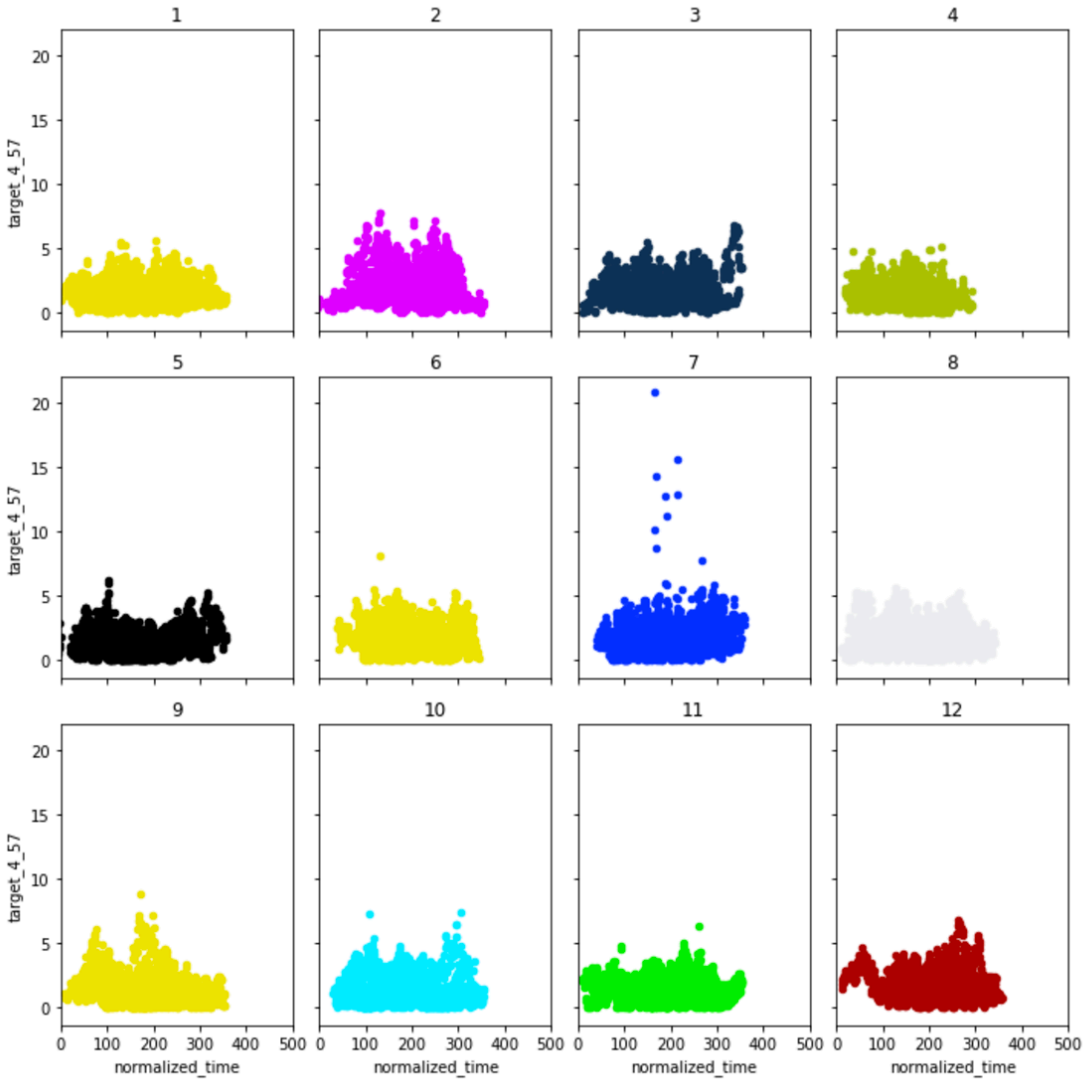


Target 4

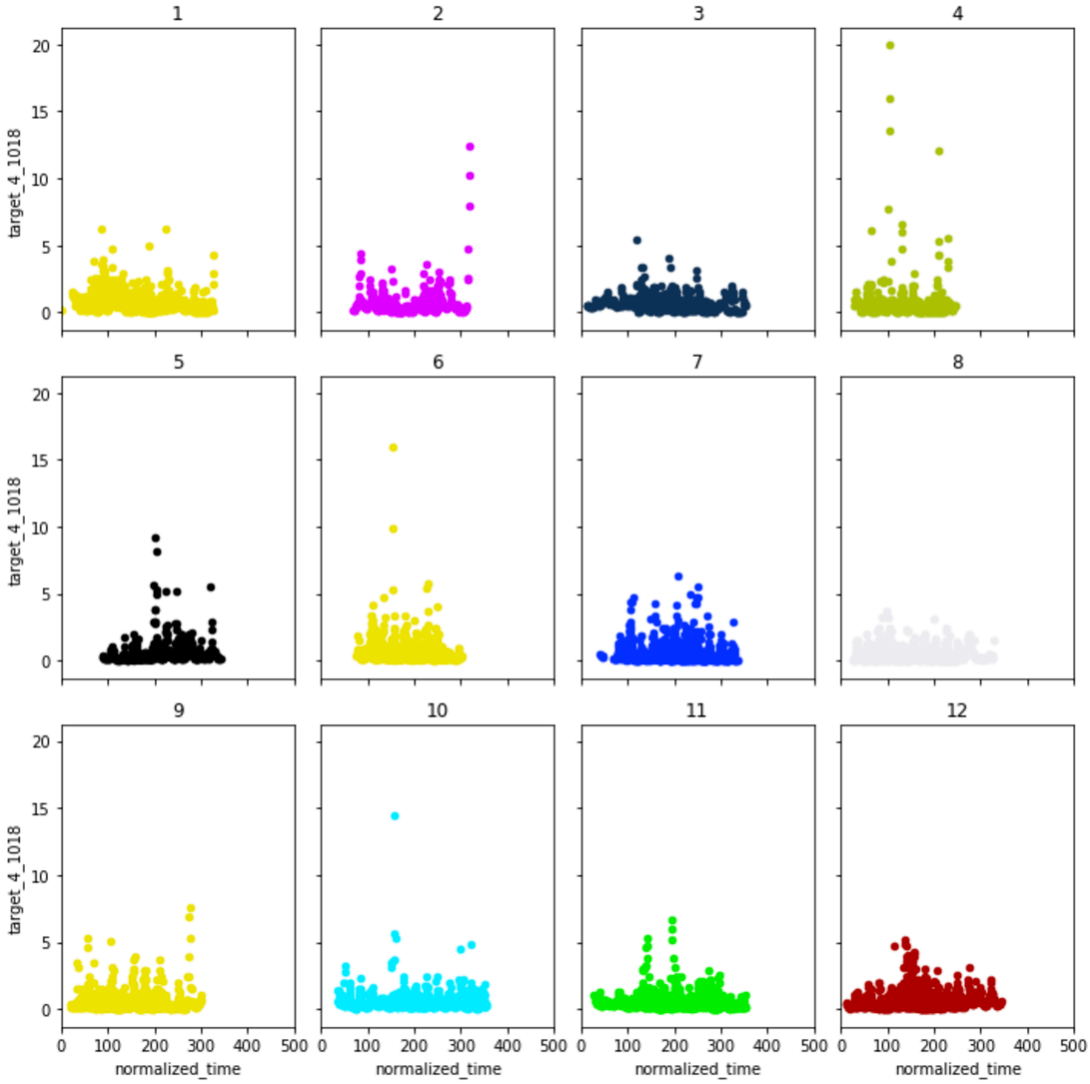
`target_4_50`



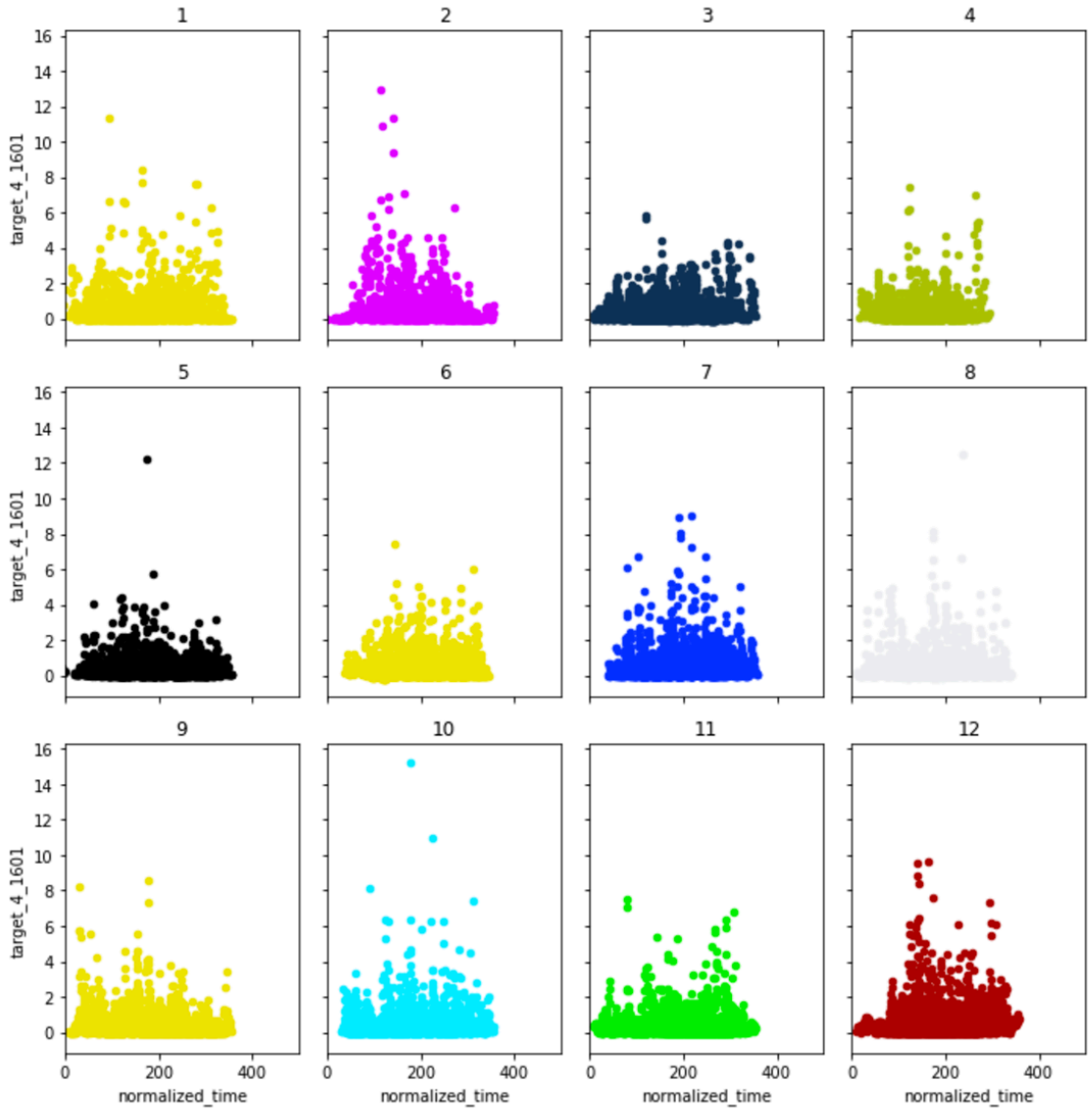
target_4_57


1
2
3
4
5
6
7
8
9
10
11
12
target_4_57
normalized_time
0
5
10
15
20
0
100
200
300
400
500

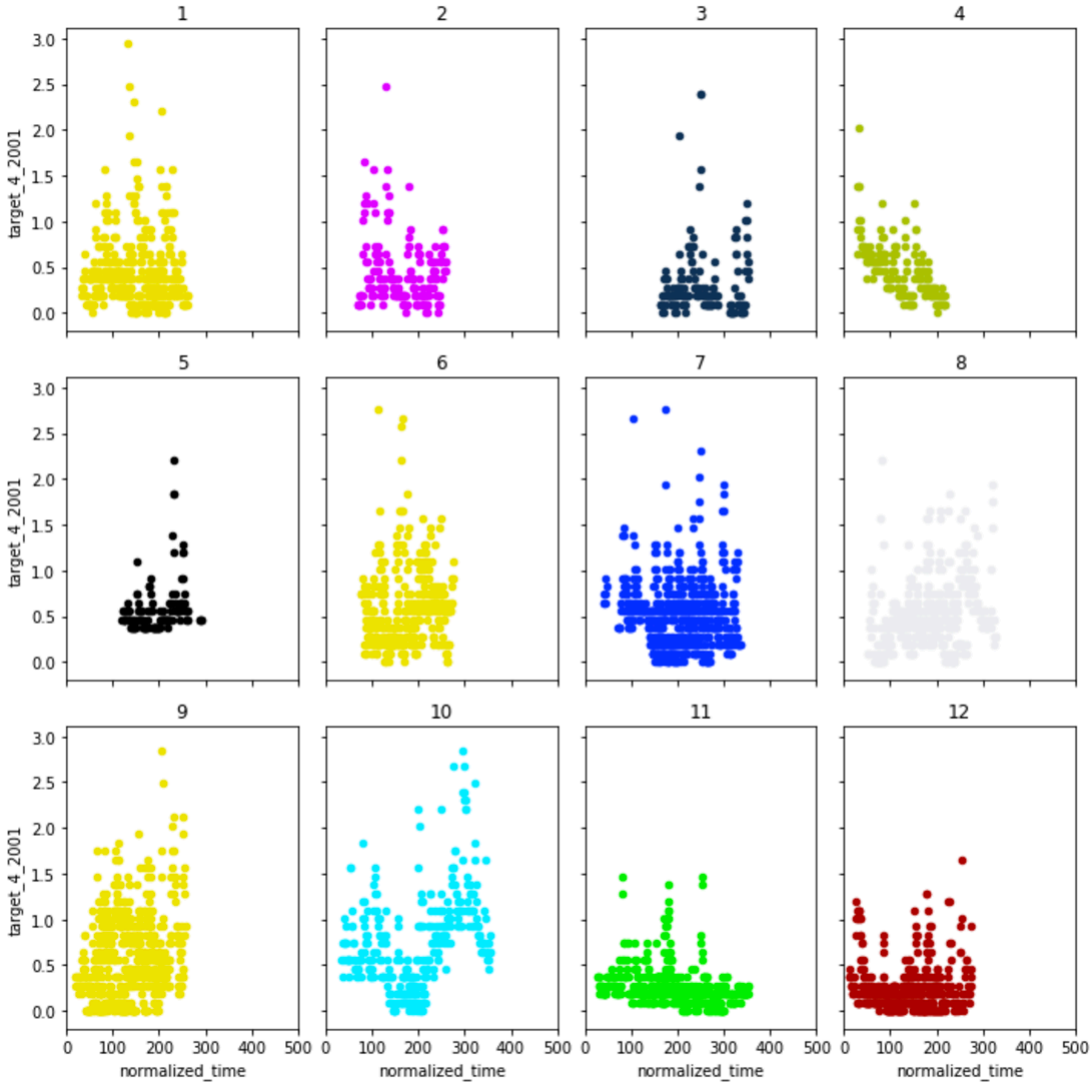
```
target_4_1018
```



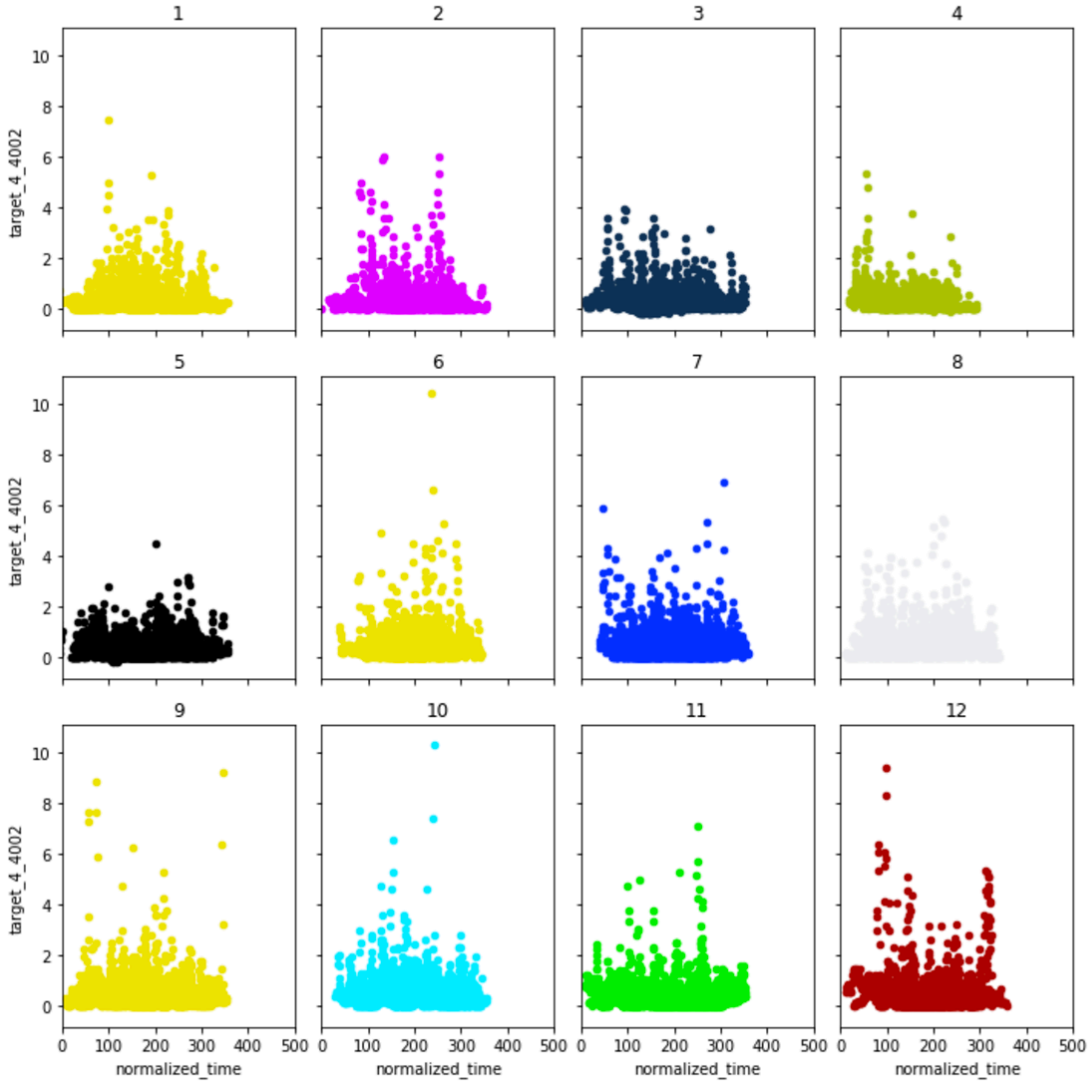
target_4_1601



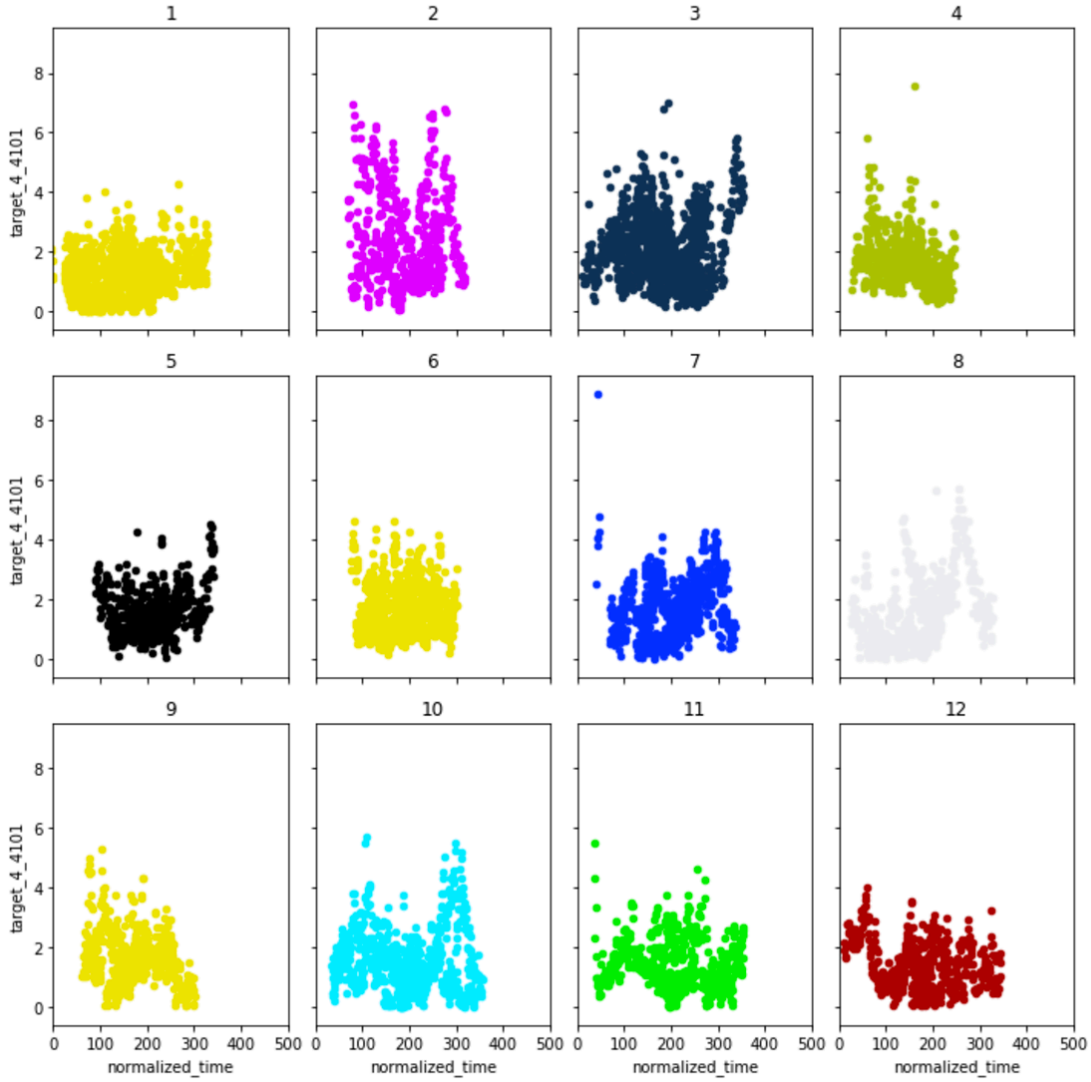
target_4_2001



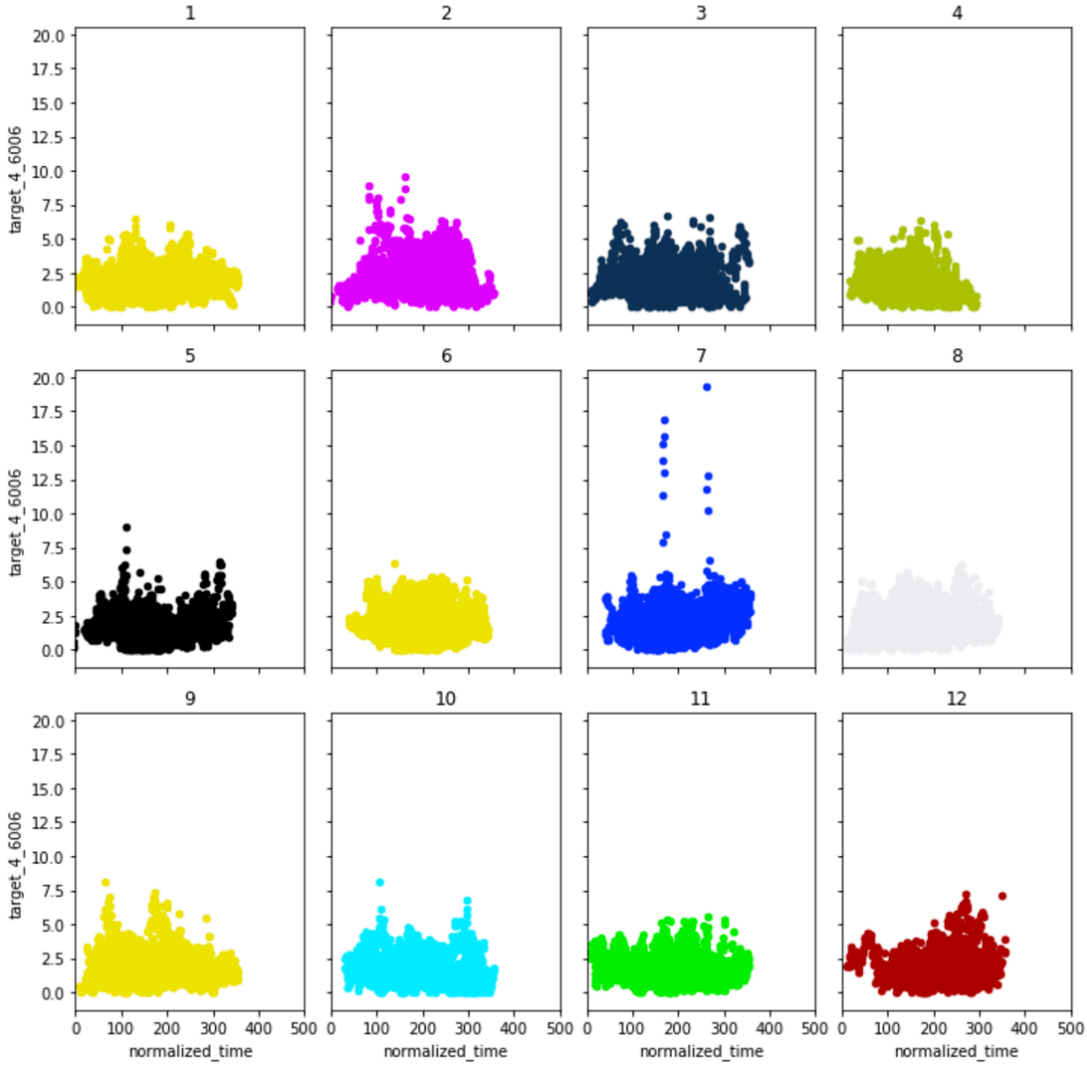
target_4_4002



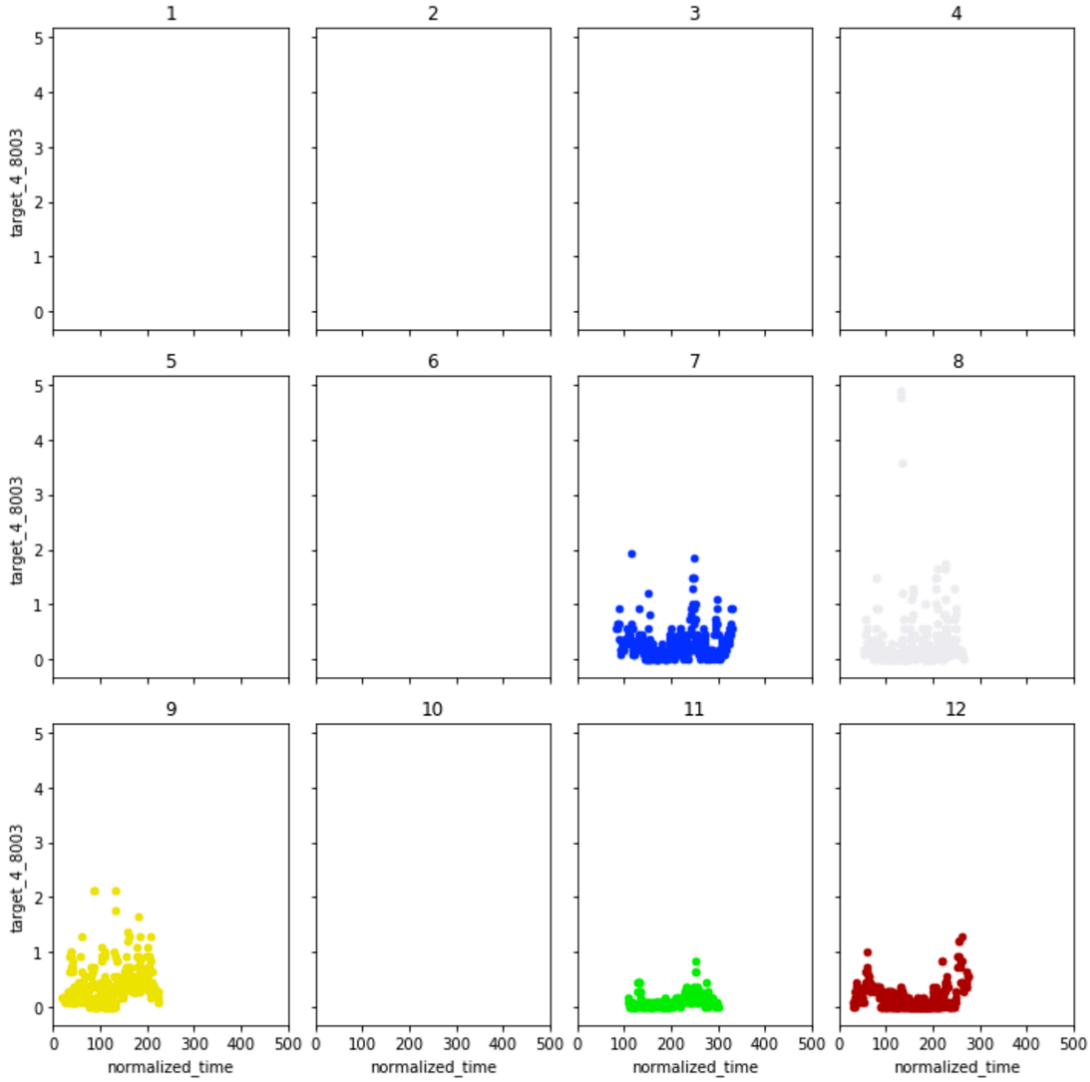
target_4_4101



target_4_6006

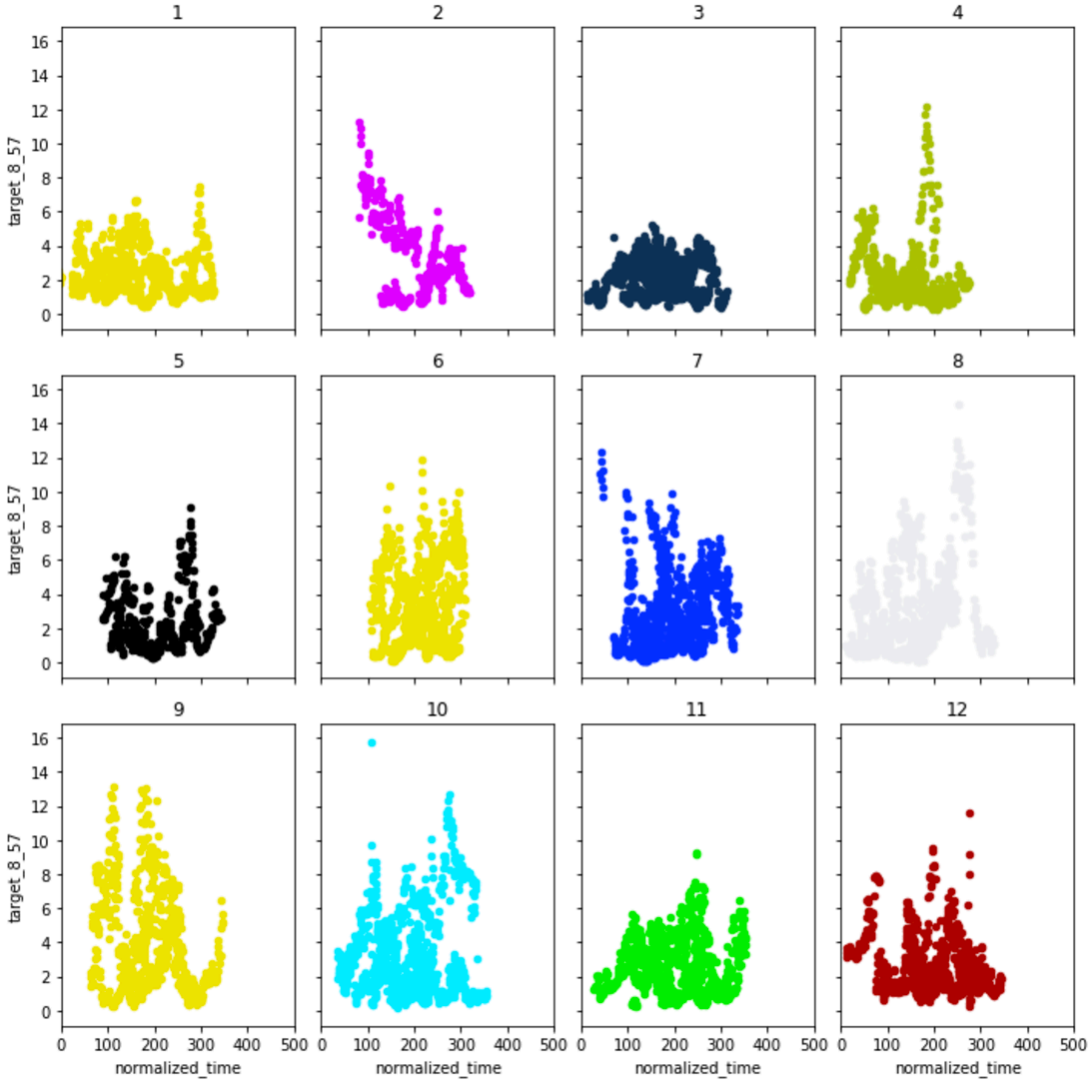


target_4_8003

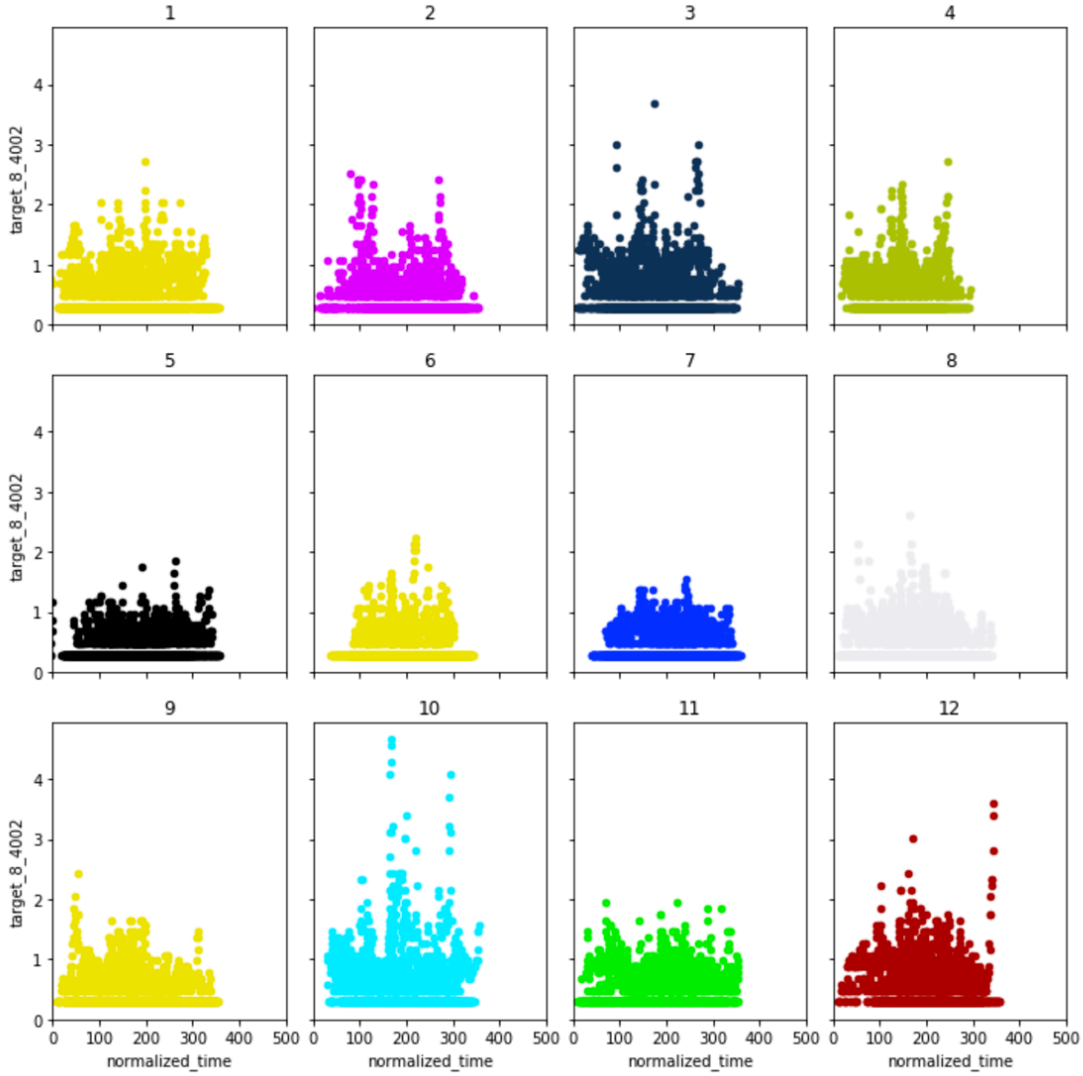


Target 8

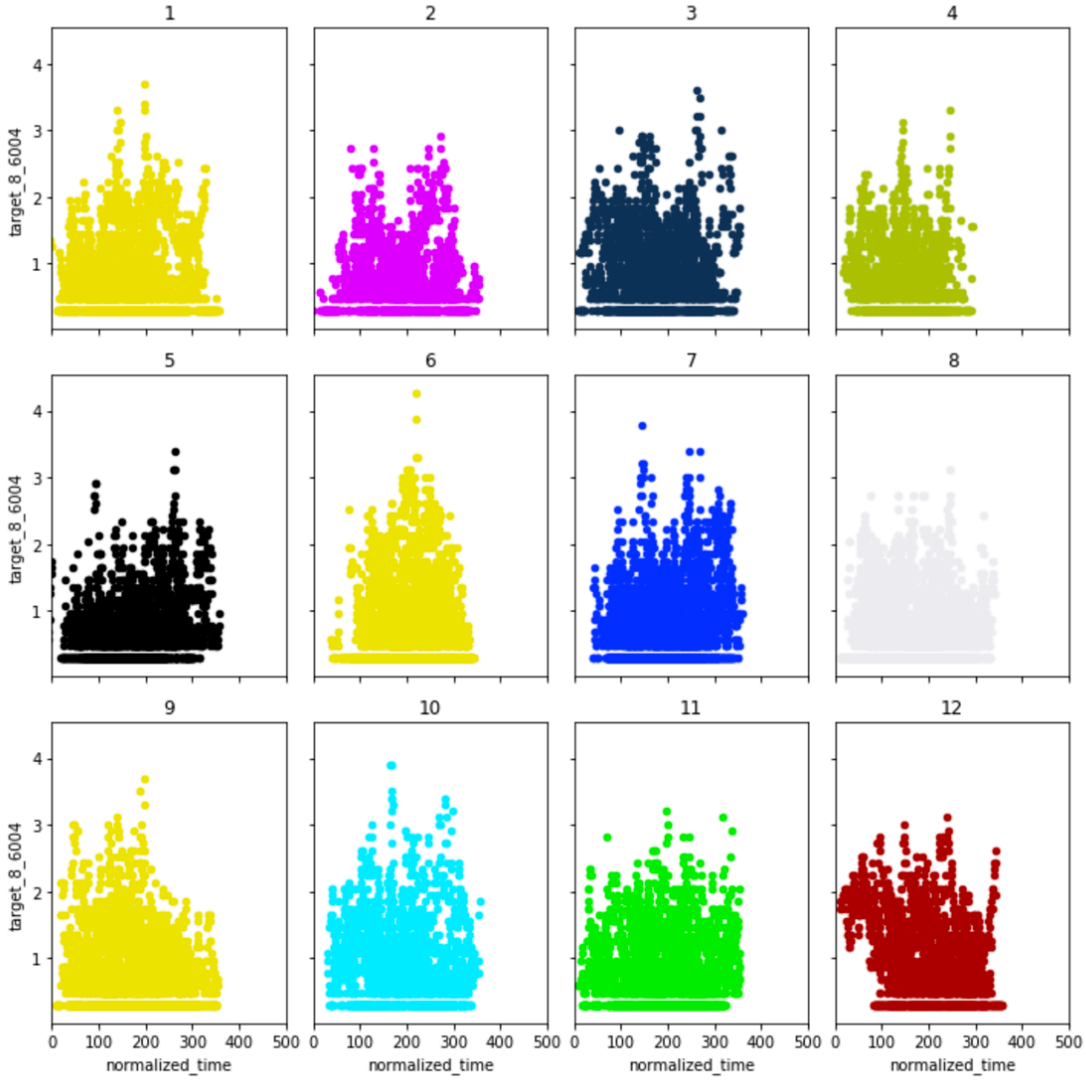
`target_8_57`



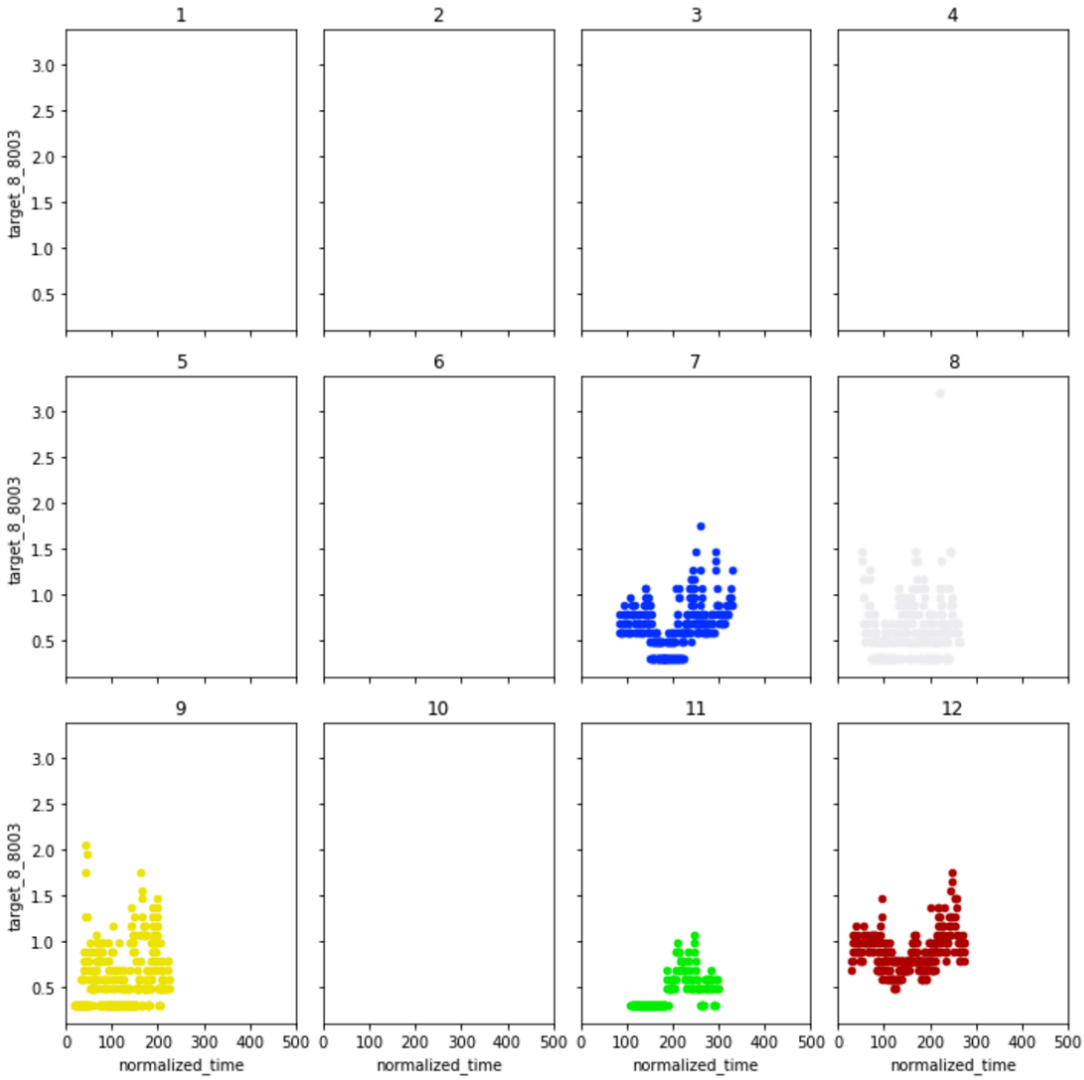
target_8_4002



target_8_6004

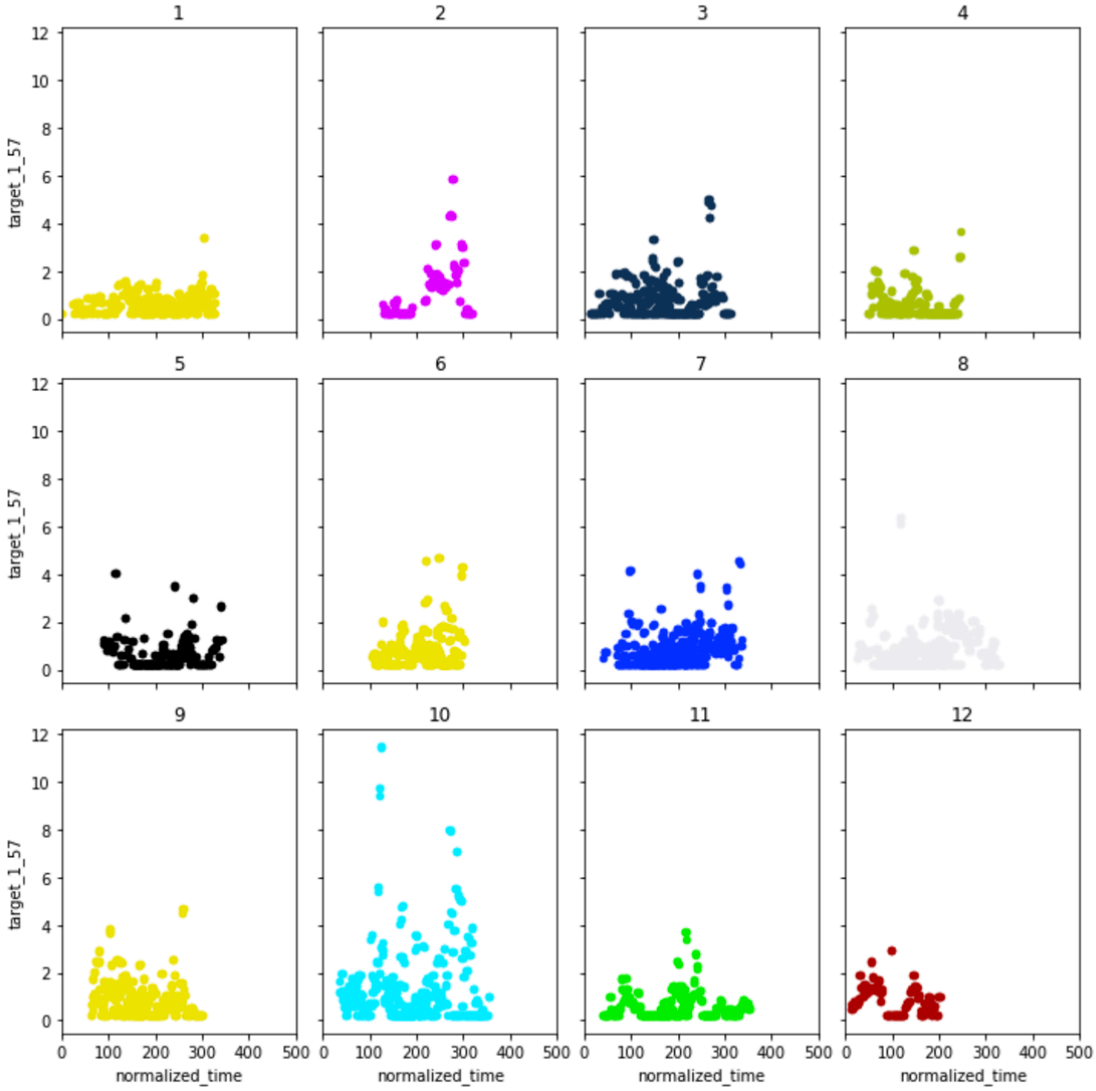


target_8_8003

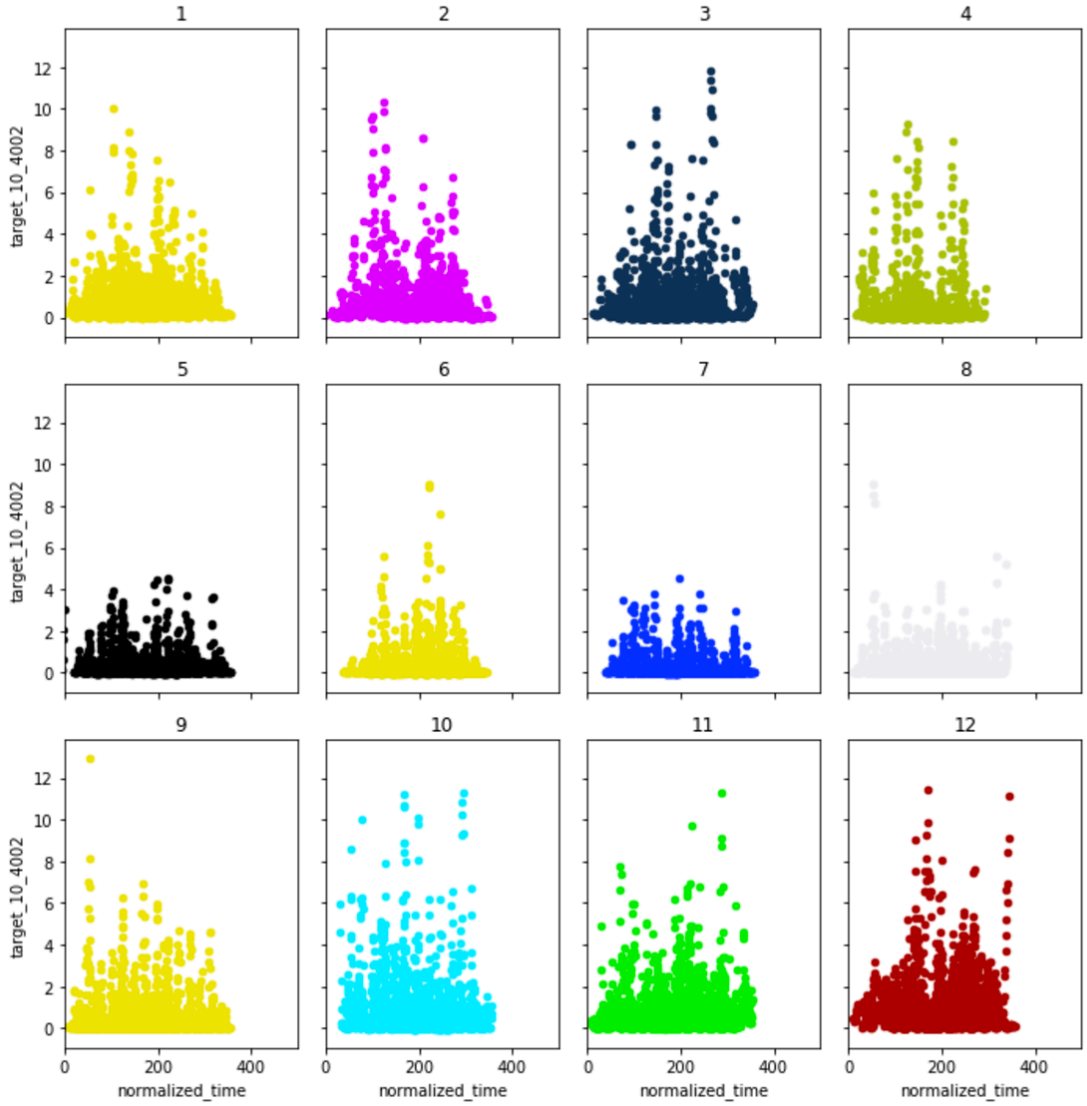


Miscellaneous Targets:

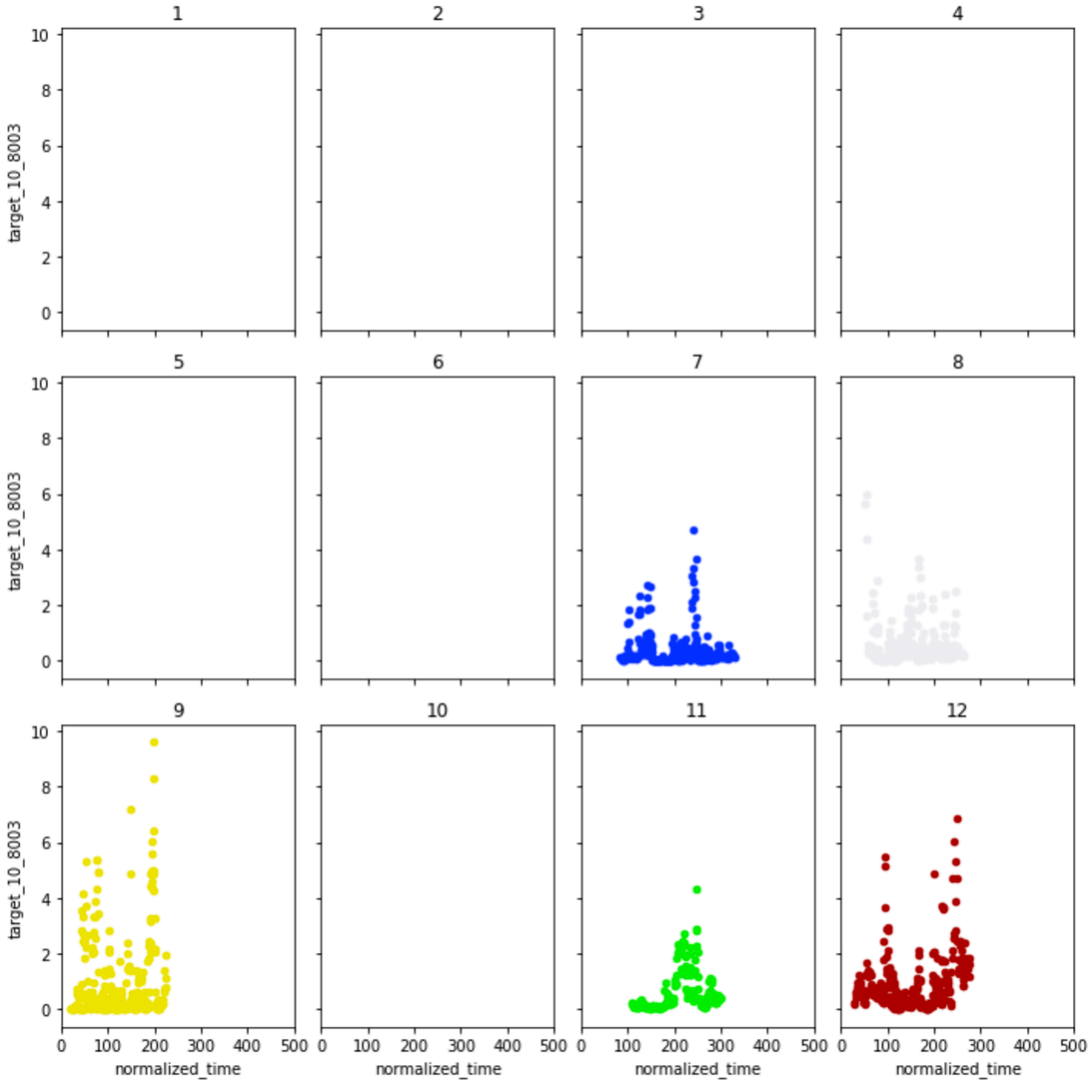
`target_1_57`



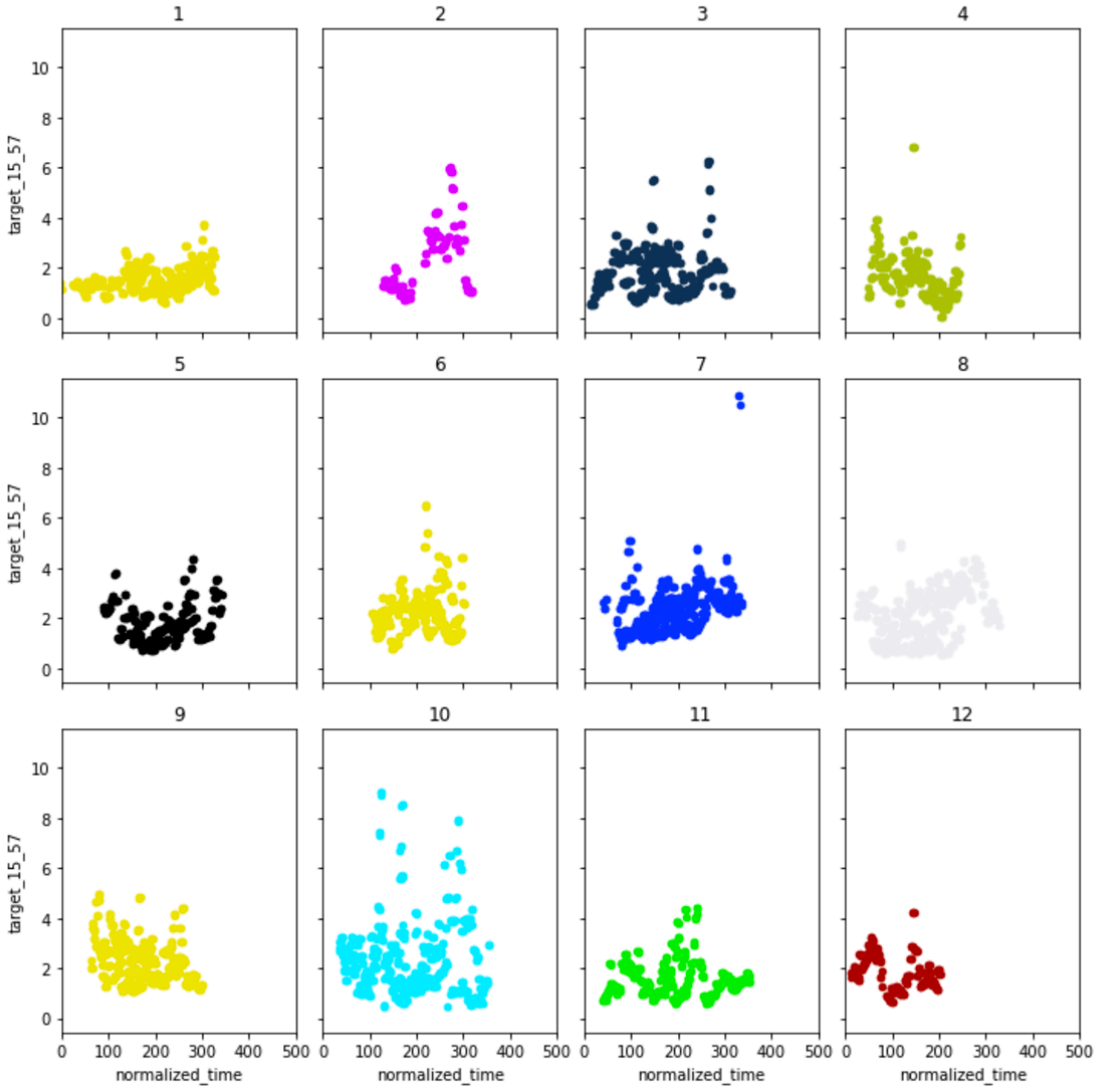
target_10_4002



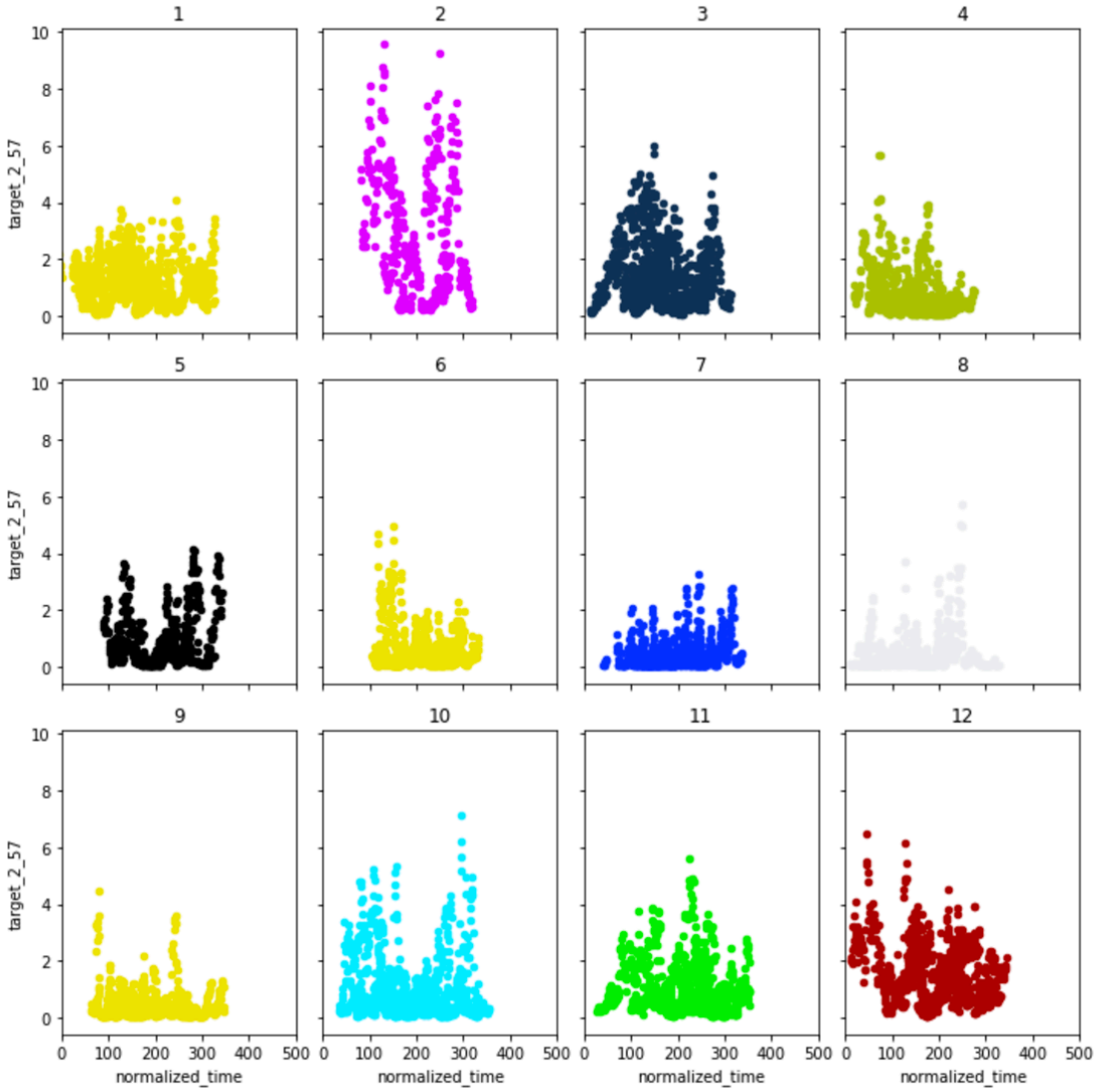
target_10_8003



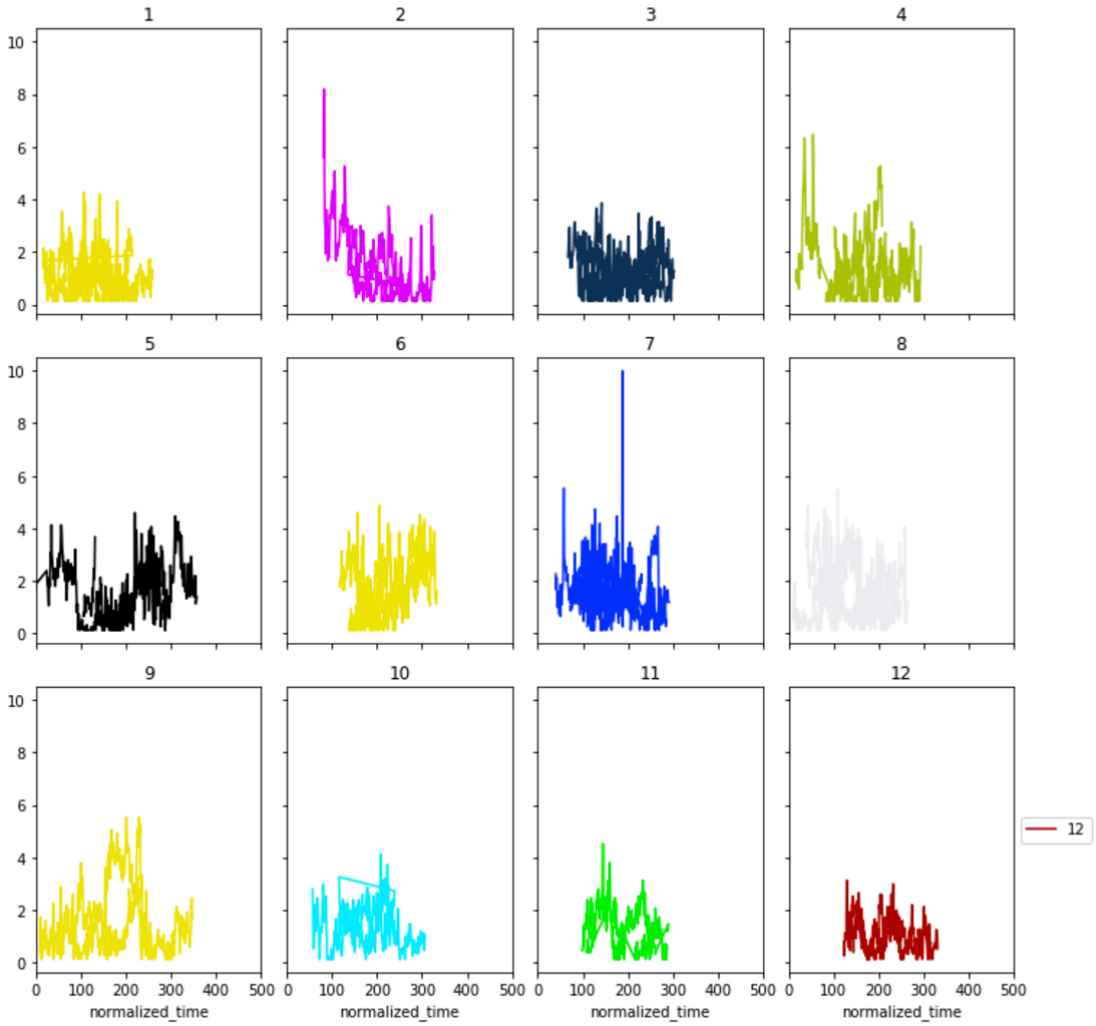
target_15_57


1
2
3
4
5
6
7
8
9
10
11
12
target_15_57
normalized_time
0
2
4
6
8
10
0
100
200
300
400
500

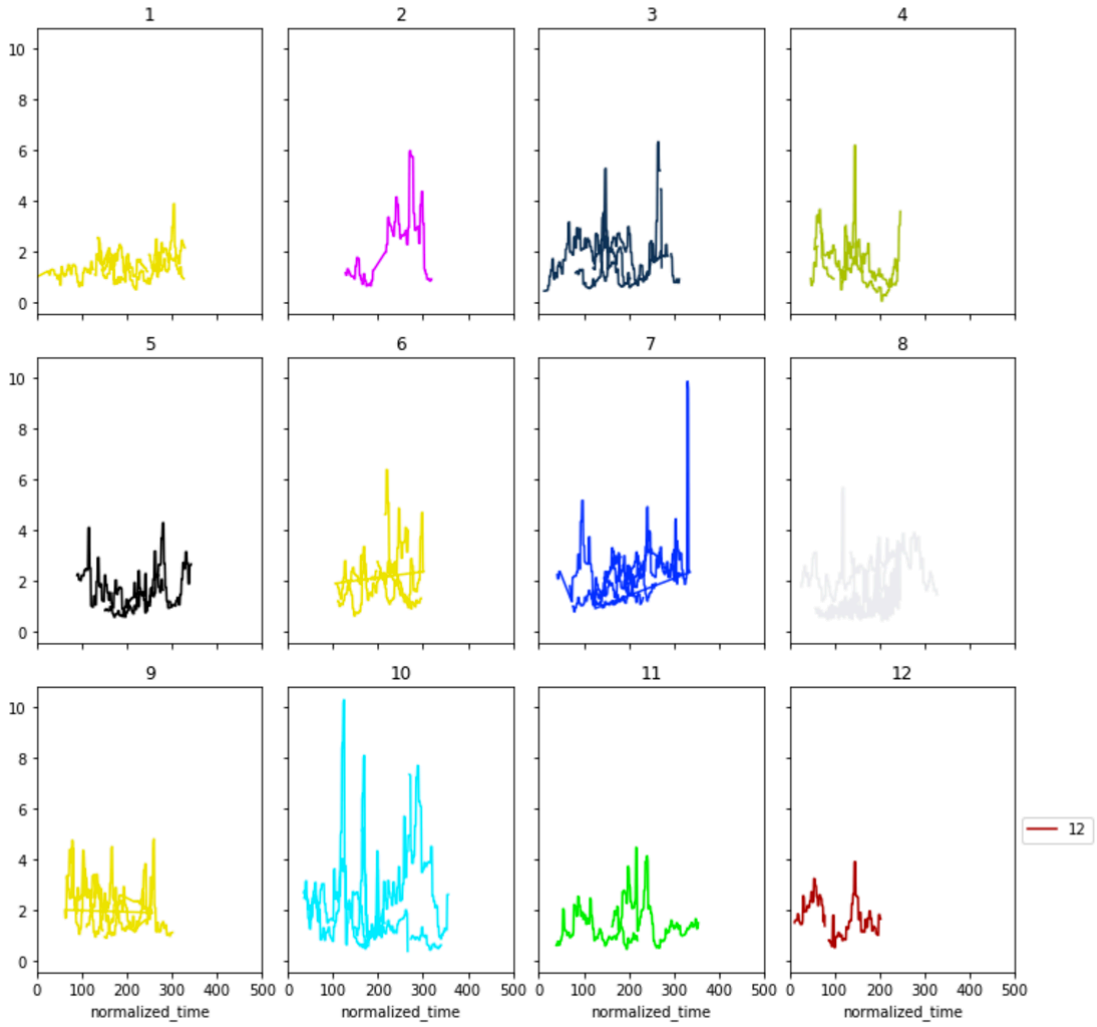
target_2_57


1
2
3
4
5
6
7
8
9
10
11
12
target_2_57
normalized_time
0
2
4
6
8
10
0
100
200
300
400
500

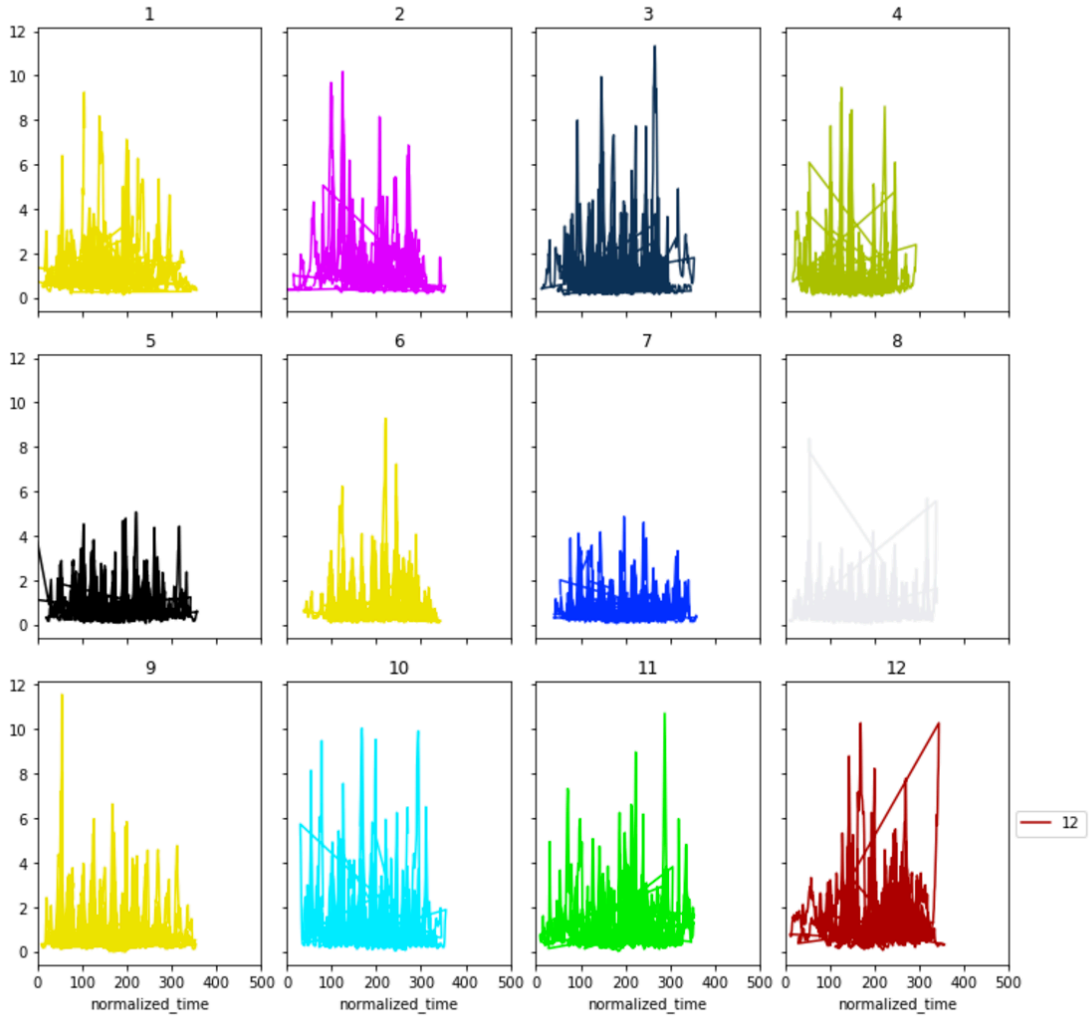
target_5_6006



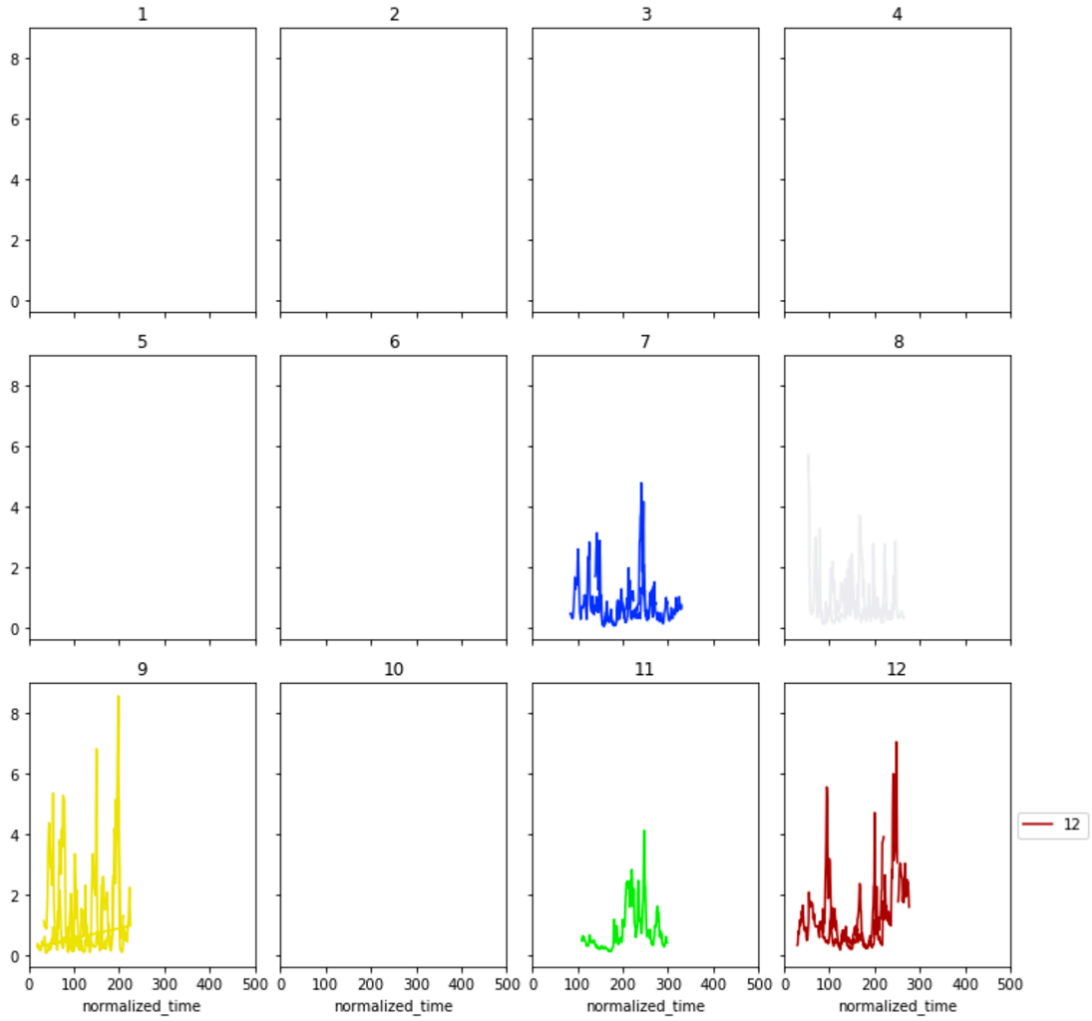
target_7_57



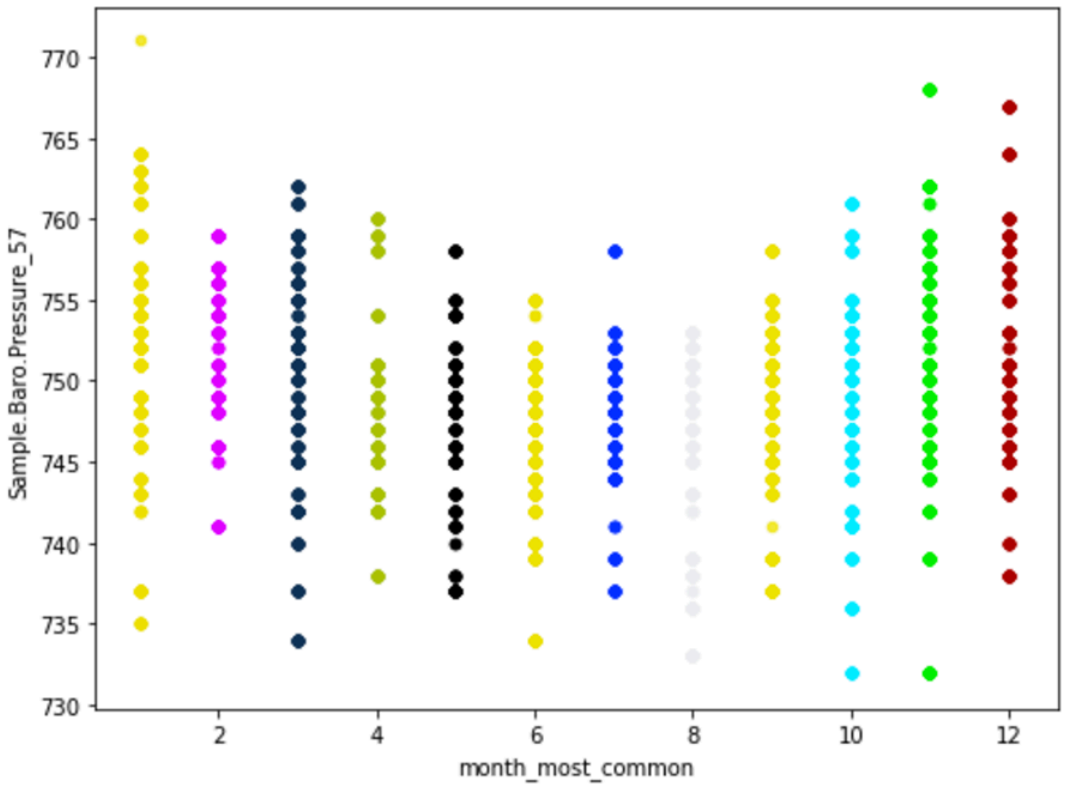
target_9_4002

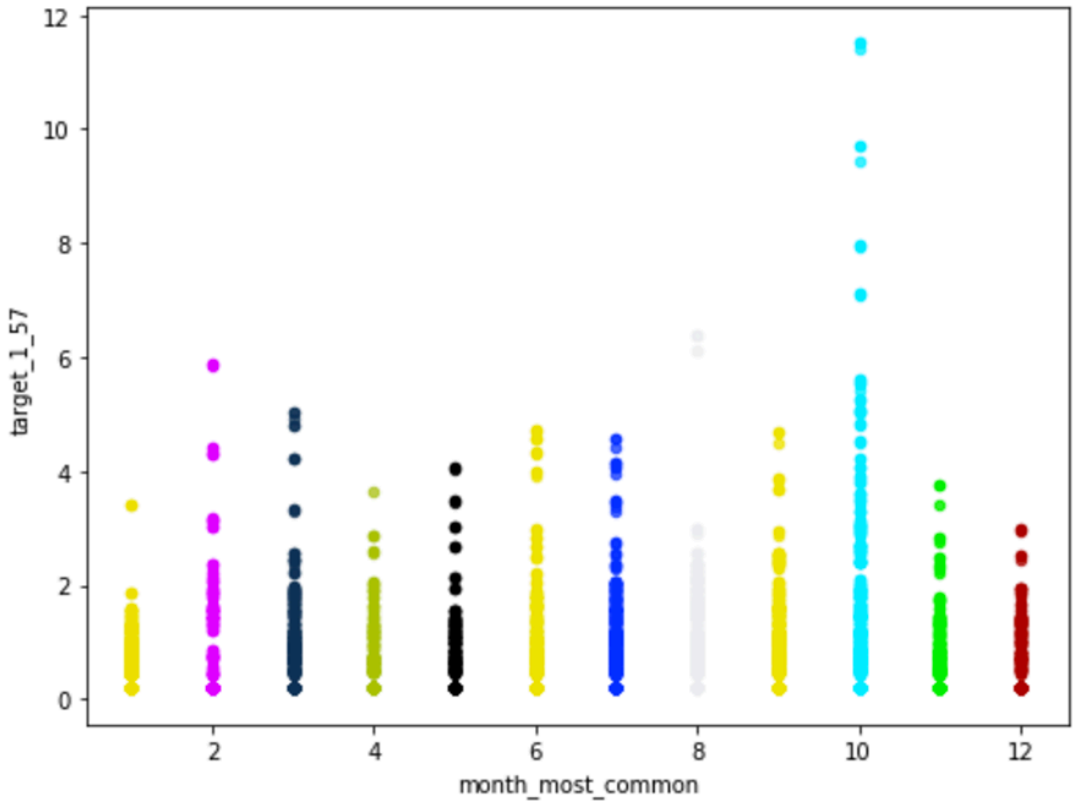


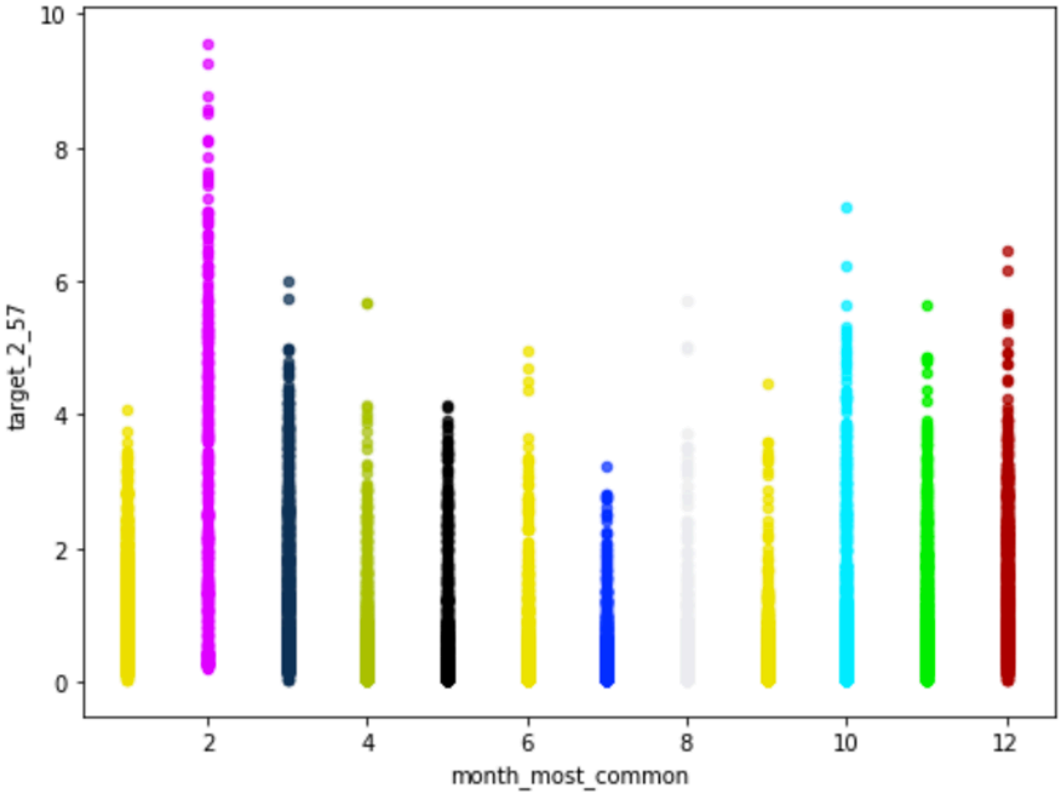
target_9_8003

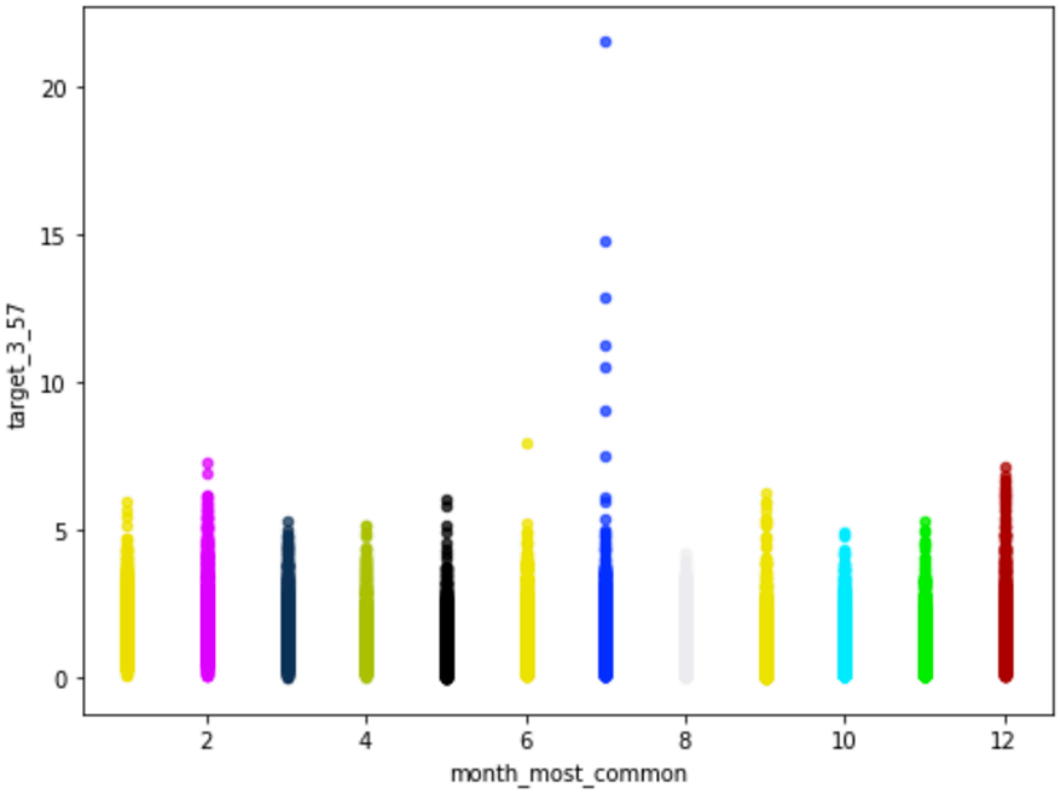


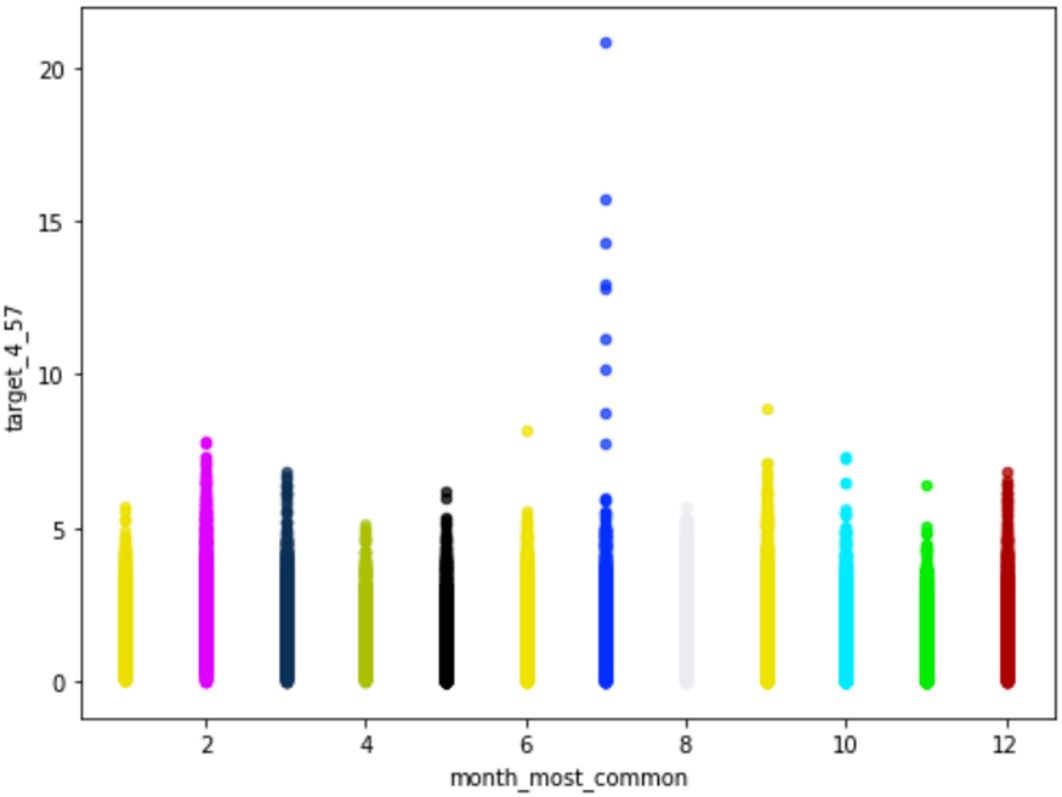
Site 57

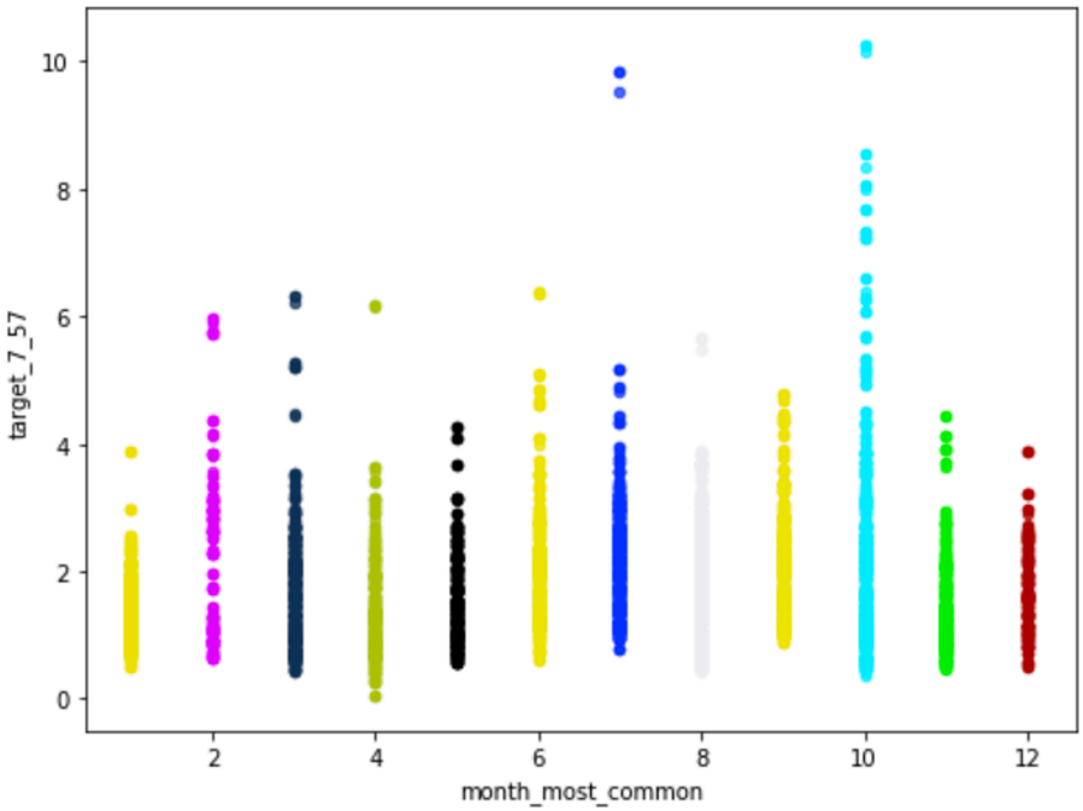

Sample.Baro.Pressure_57
770
765
760
755
750
745
740
735
730
2
4
6
8
10
12
month_most_common

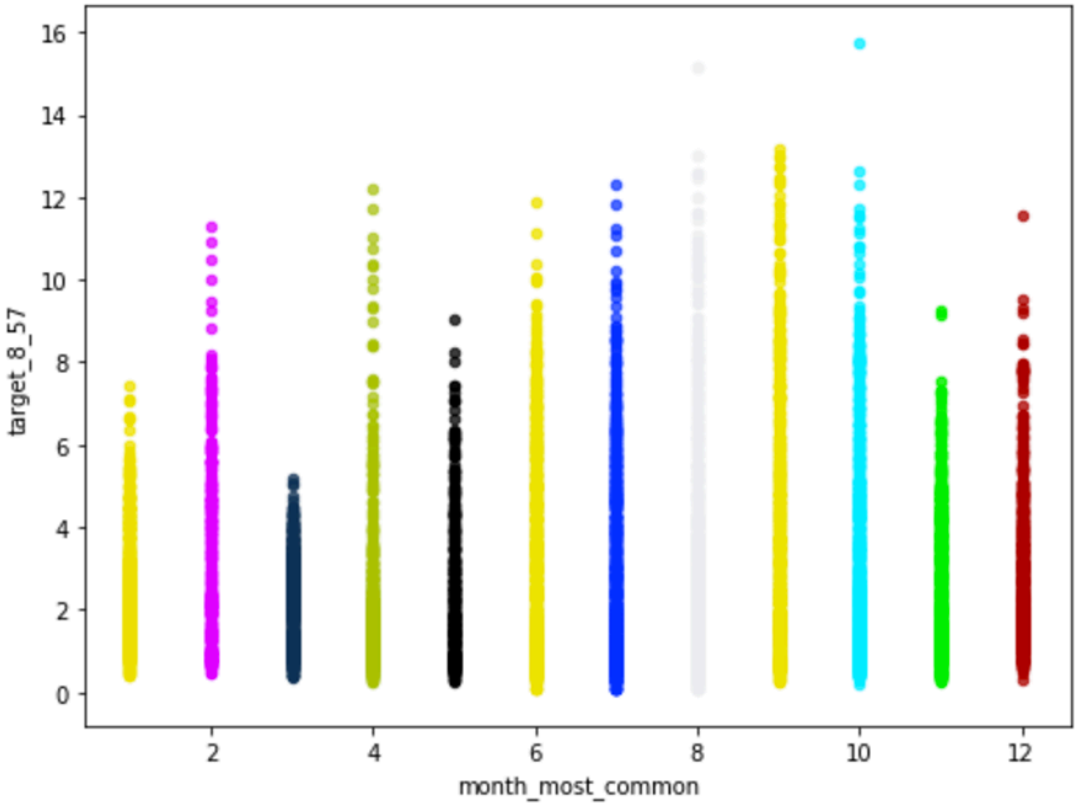

target_1_57
month_most_common
0
2
4
6
8
10
12

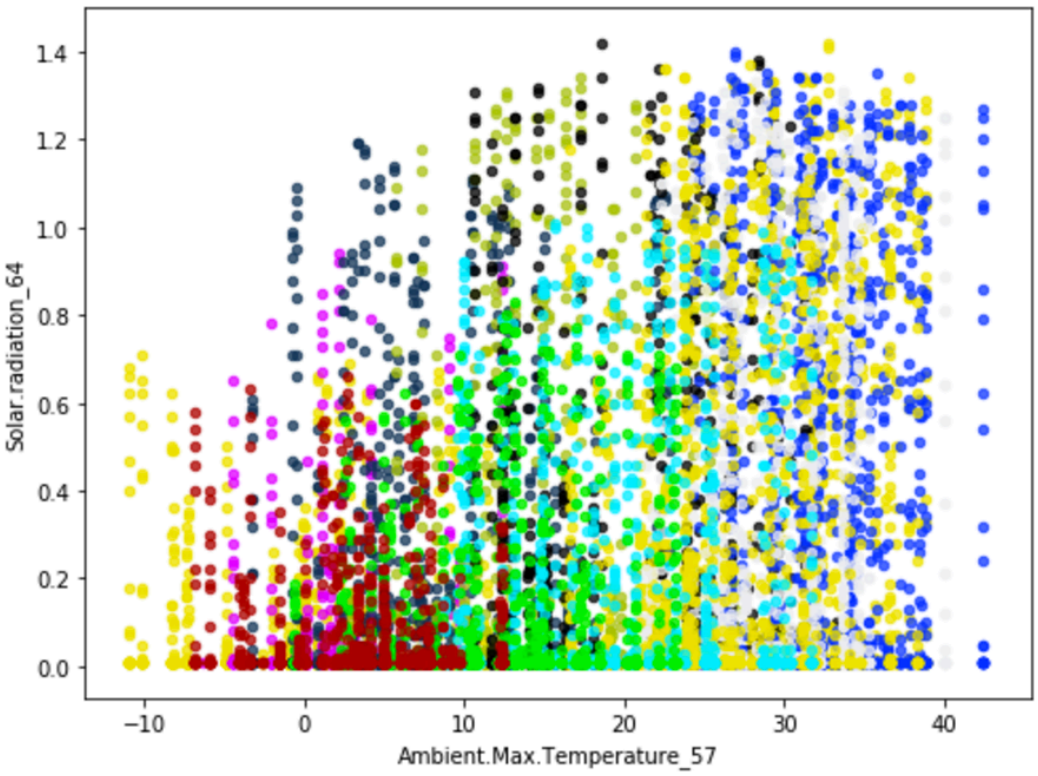

target_2_57
month_most_common
0
2
4
6
8
10

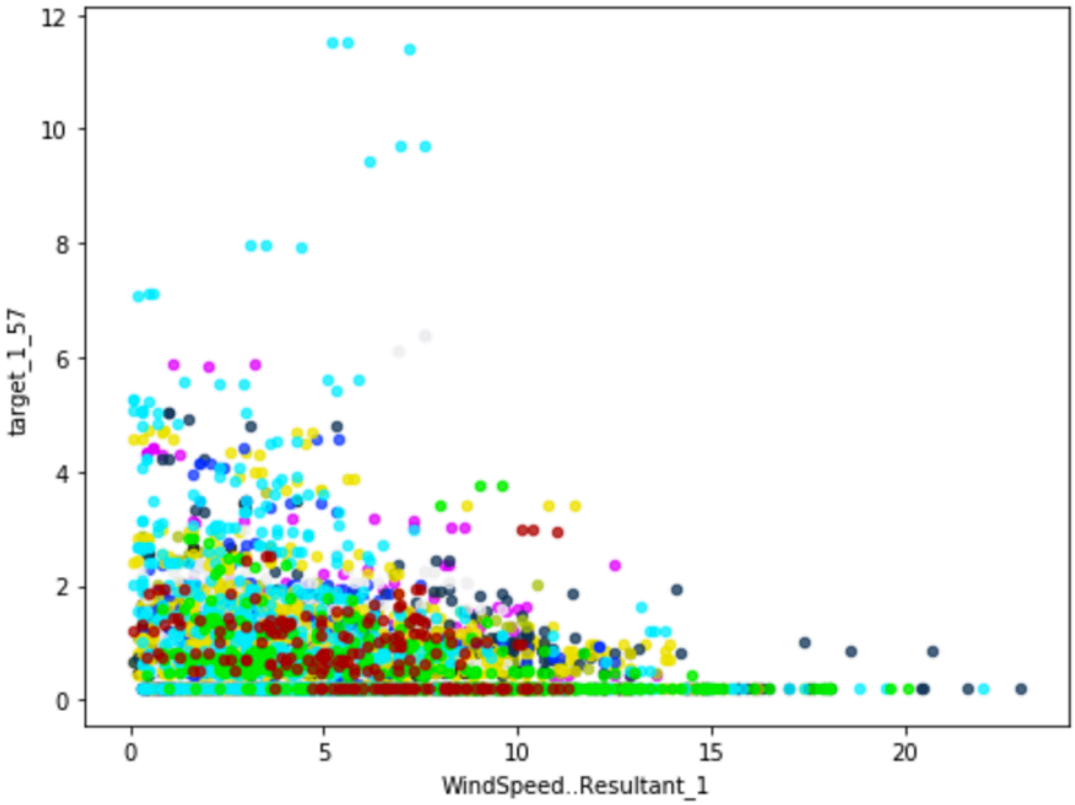

target_3_57
20
15
10
5
0
2
4
6
8
10
12
month_most_common

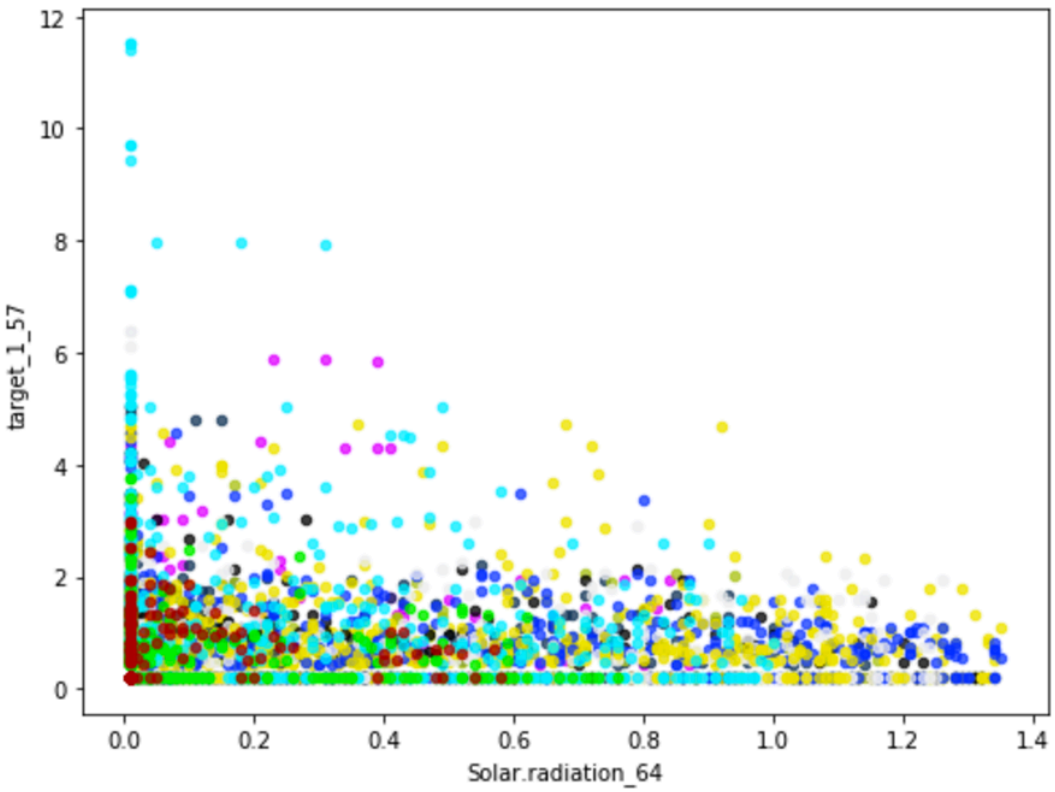

target_4_57
20
15
10
5
0
2
4
6
8
10
12
month_most_common

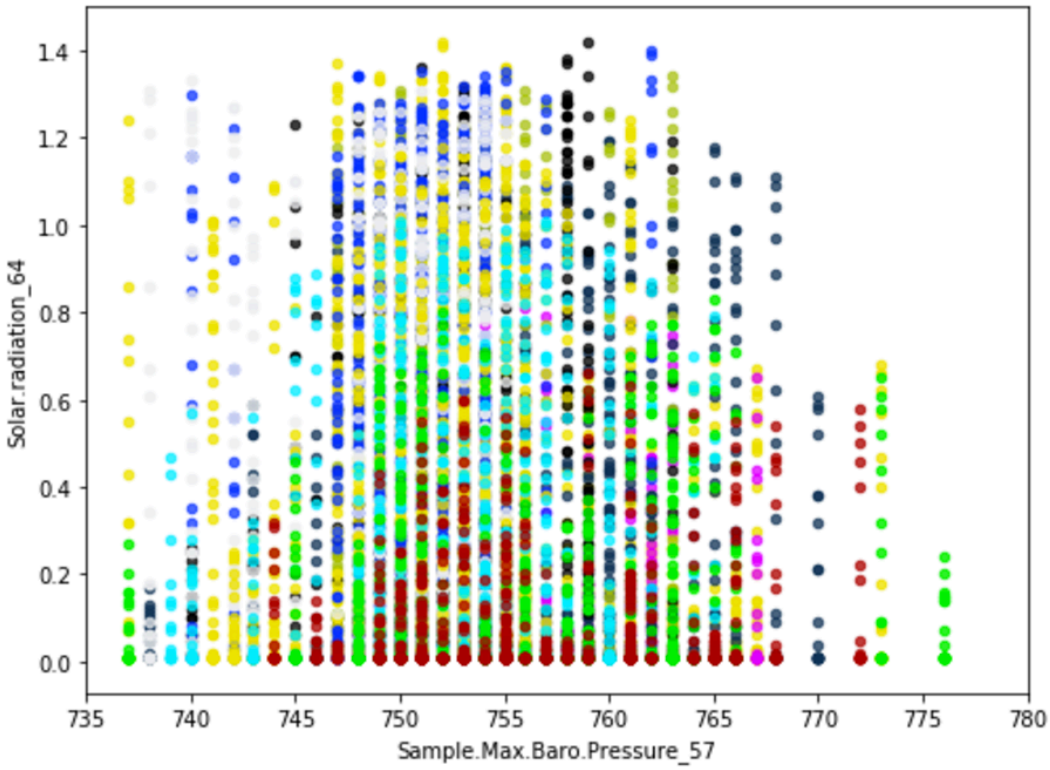

target_7_57
month_most_common
0
2
4
6
8
10
2
4
6
8
10
12

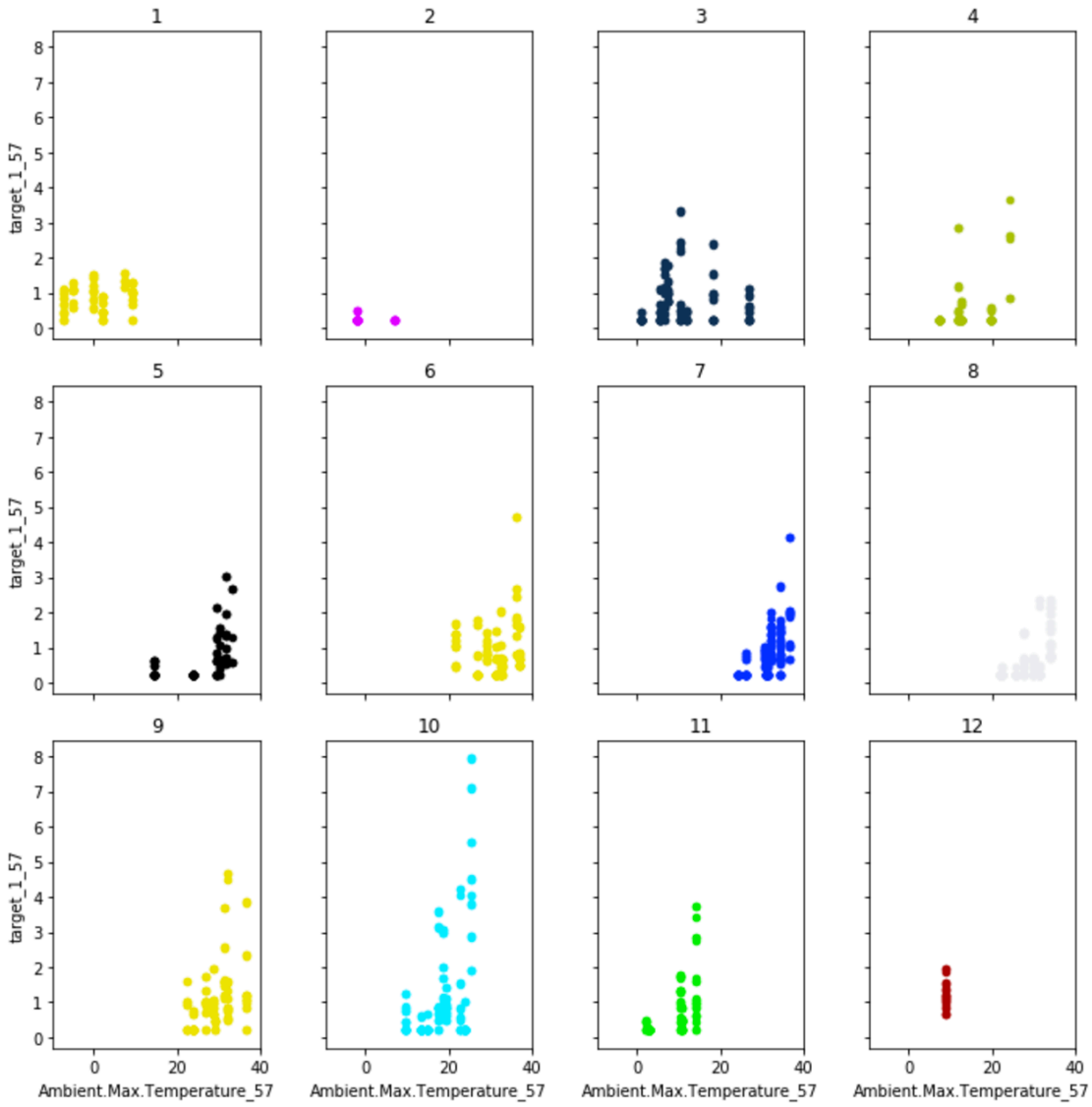

target_8_57
month_most_common
0
2
4
6
8
10
12
14
16

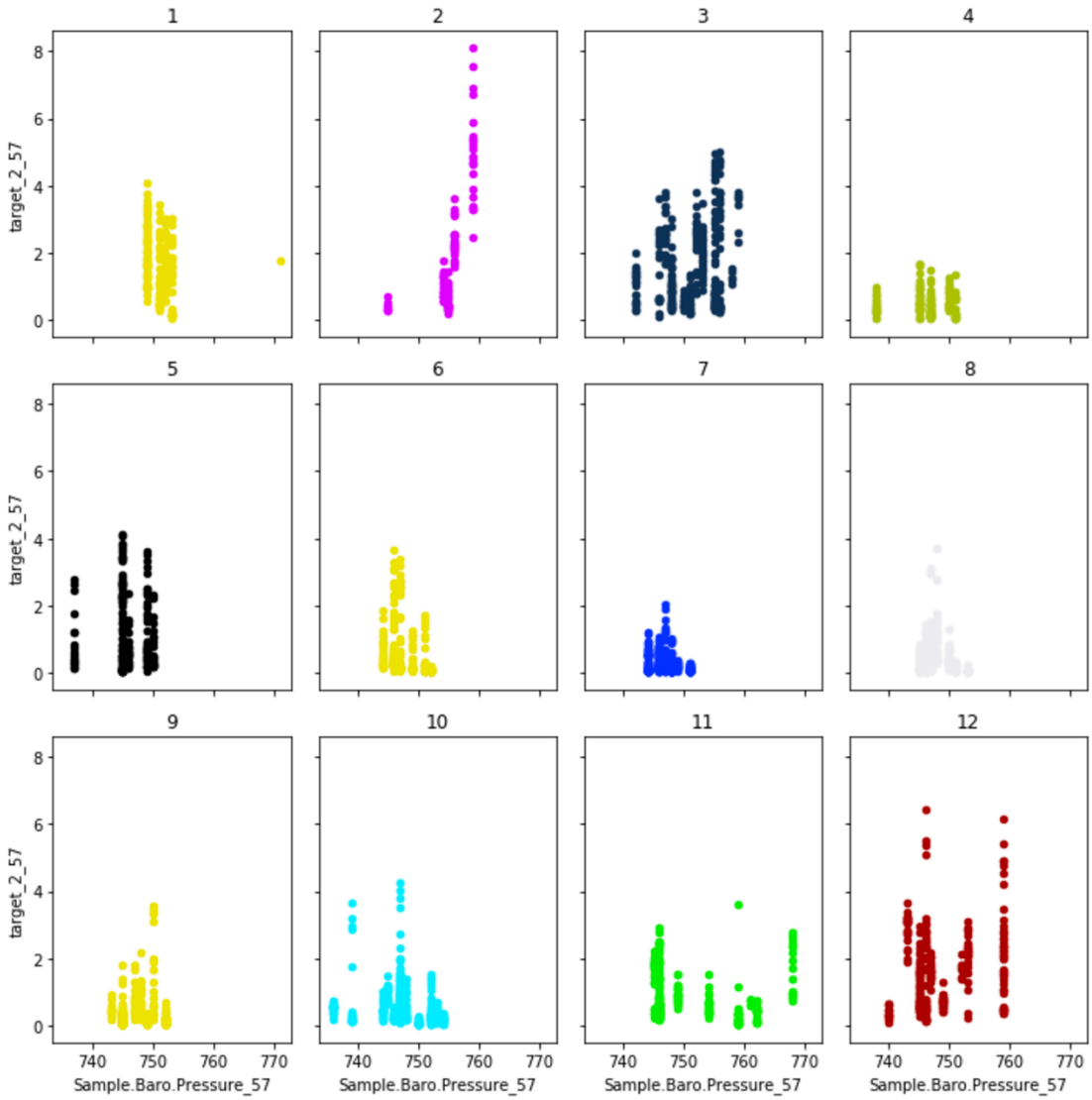

Solar.radiation_64
Ambient.Max.Temperature_57
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
−10
0
10
20
30
40

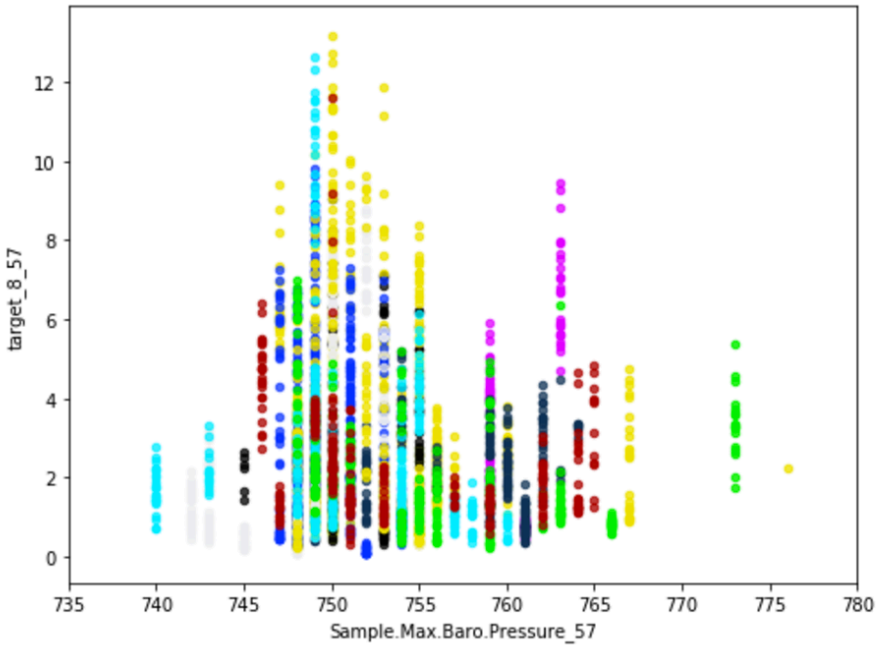

target_1_57
WindSpeed..Resultant_1
0
2
4
6
8
10
12
0
5
10
15
20

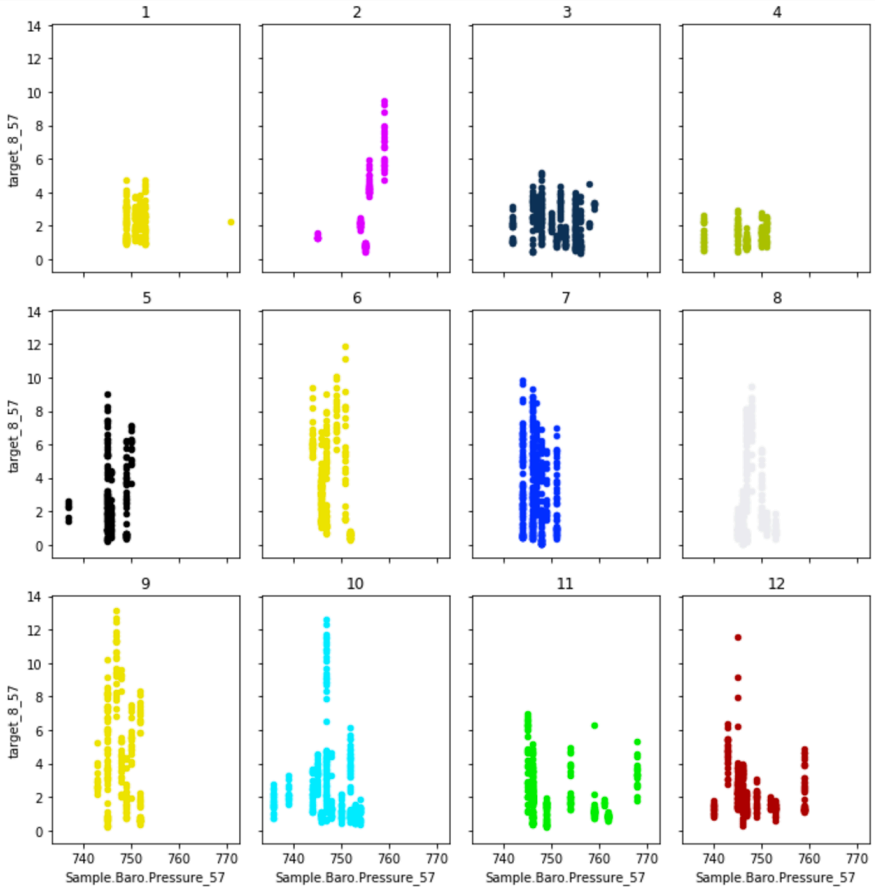

target_1_57
Solar.radiation_64
0
2
4
6
8
10
12
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4

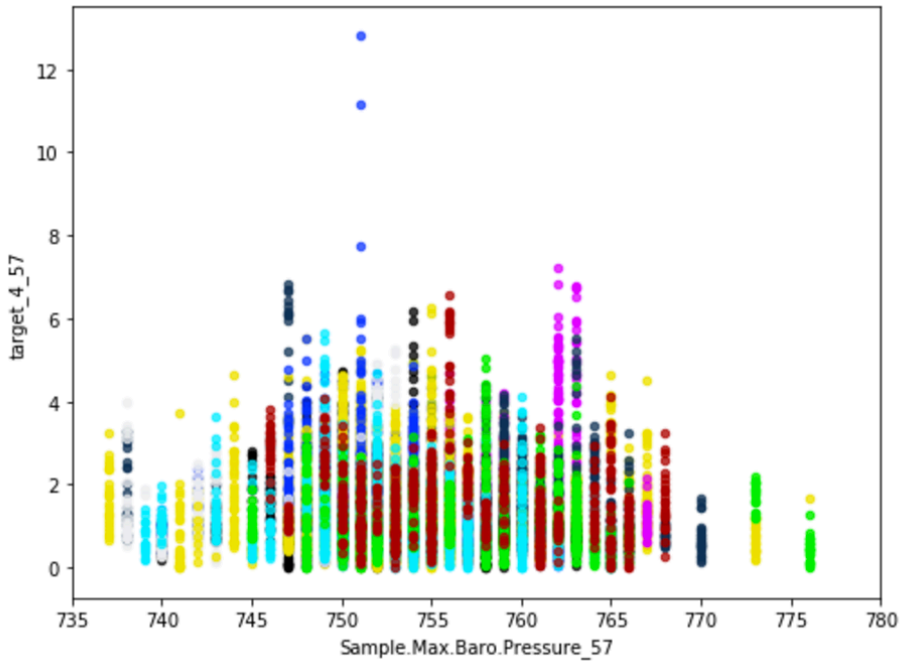

Solar.radiation_64
Sample.Max.Baro.Pressure_57
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
735
740
745
750
755
760
765
770
775
780

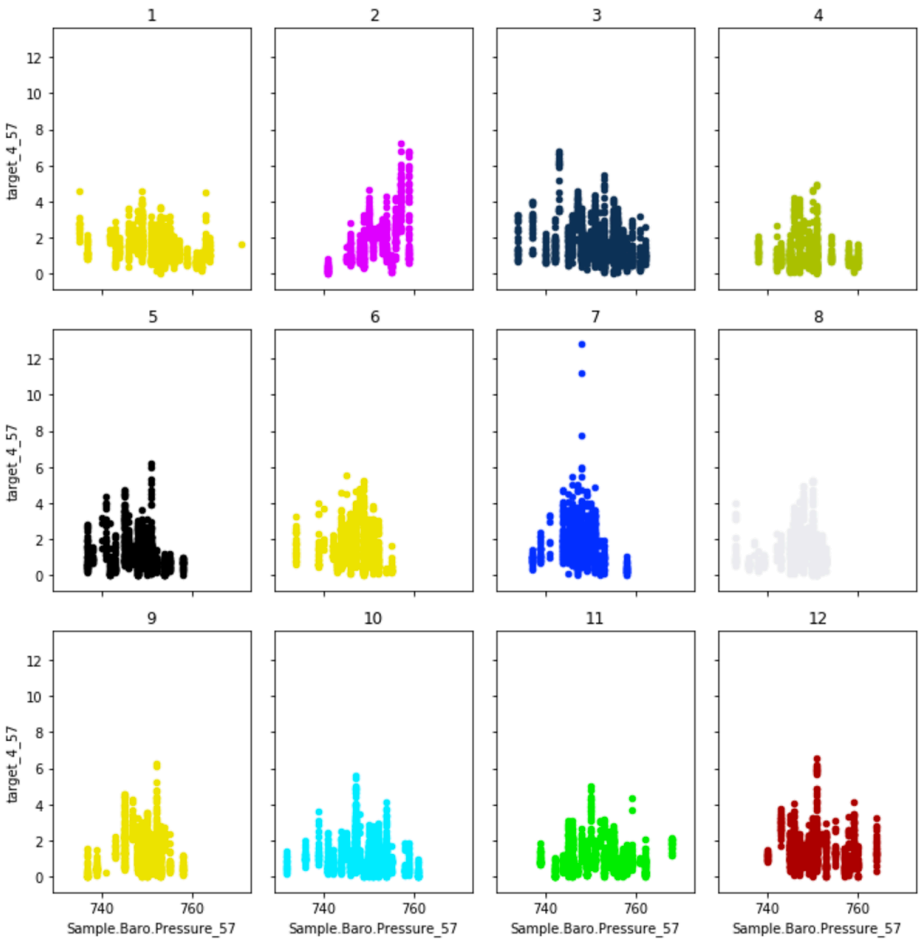

1
2
3
4
5
6
7
8
9
10
11
12
target_1_57
Ambient.Max.Temperature_57


1
2
3
4
5
6
7
8
9
10
11
12
target_2_57
Sample.Baro.Pressure_57
0
2
4
6
8
740
750
760
770


target_8_57
Sample.Max.Baro.Pressure_57

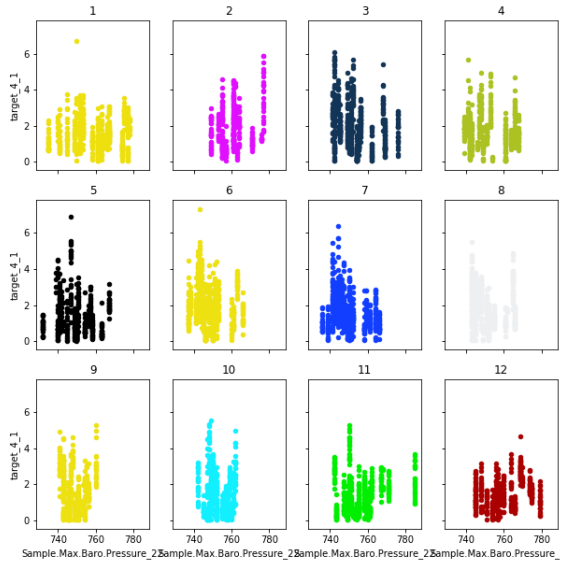

1
2
3
4
5
6
7
8
9
10
11
12
target_8_57
Sample.Baro.Pressure_57

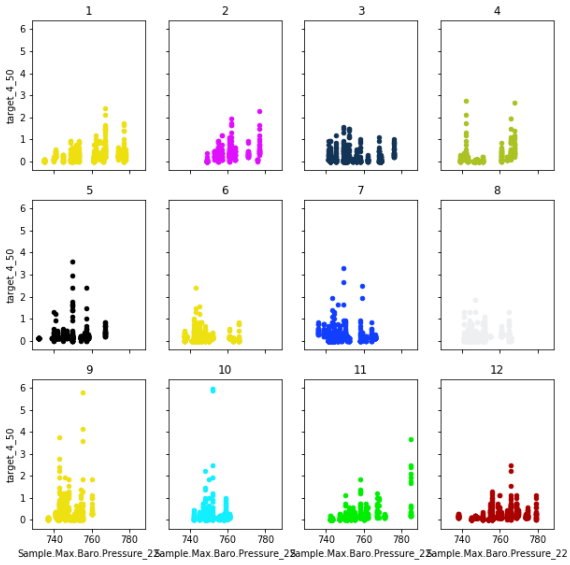

target_4_57
Sample.Max.Baro.Pressure_57
0
2
4
6
8
10
12
735
740
745
750
755
760
765
770
775
780

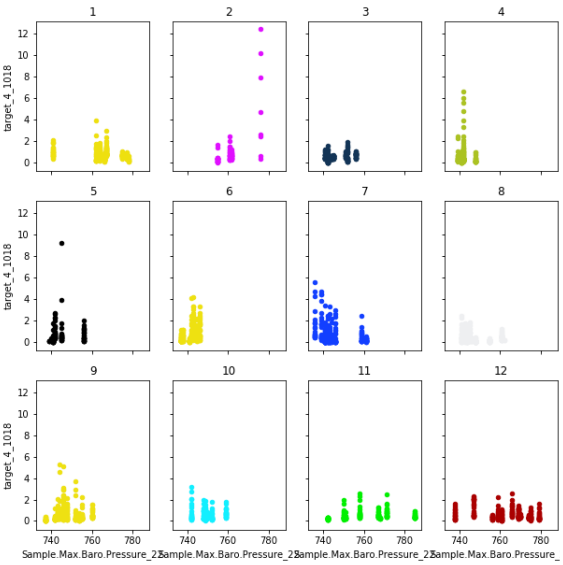

1
2
3
4
5
6
7
8
9
10
11
12
target_4_57
Sample.Baro.Pressure_57
0
2
4
6
8
10
12
740
760

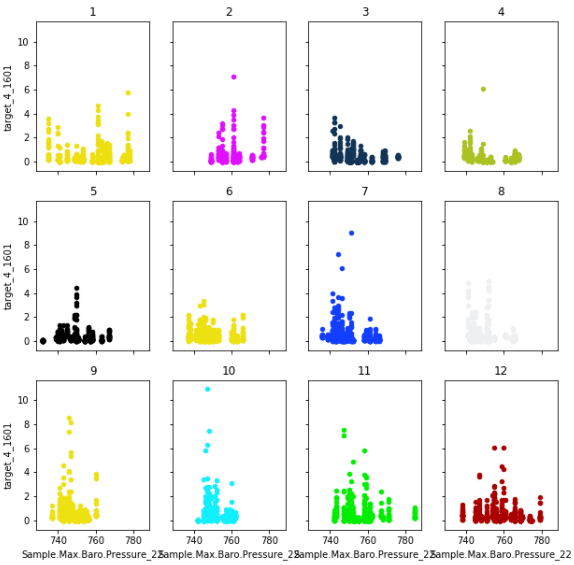
I. Target 4
    I. Barometric Pressure
    II. Solar Radiation
    III. Temperature
II. Target 11
    I. Barometric Pressure
    II. Solar Radiation
    III. Temperature
III. Site 4002
    I. Solar Radiation
    II. Temperature
    III. Windspeed
IV. Site 57
    I. Barometric Pressure
    II. Solar Radiation
    III. Temperature
    IV. Windspeed
V. Site 8003

Target 4 vs. Barometric Pressure 1

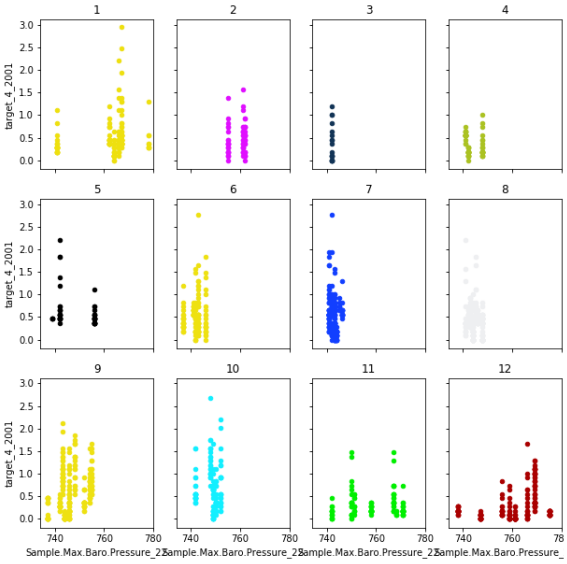


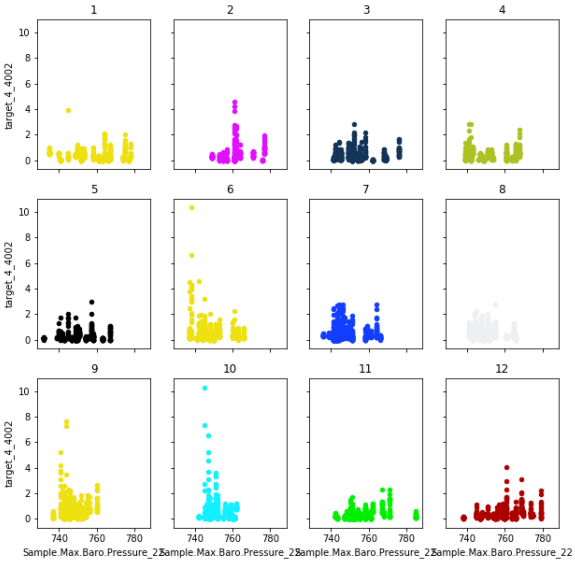


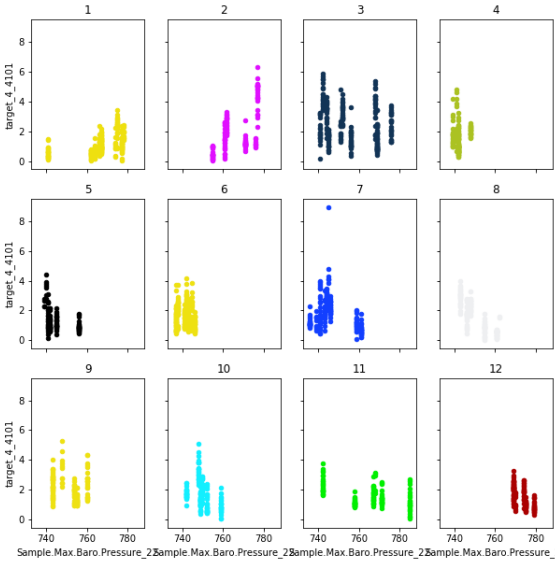


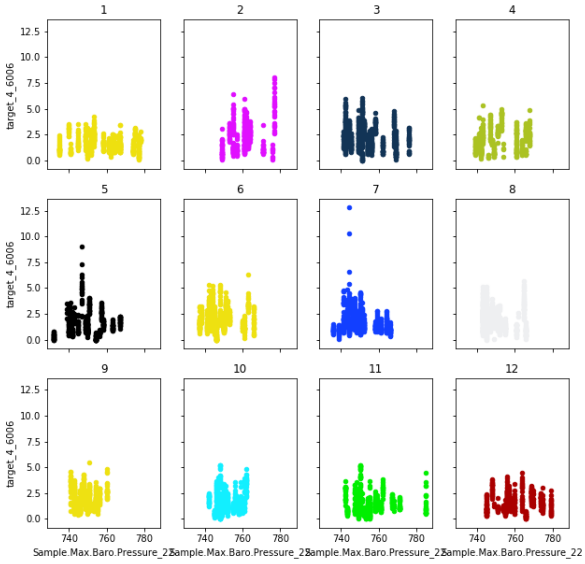


Target 4 vs. Barometric Pressure 2

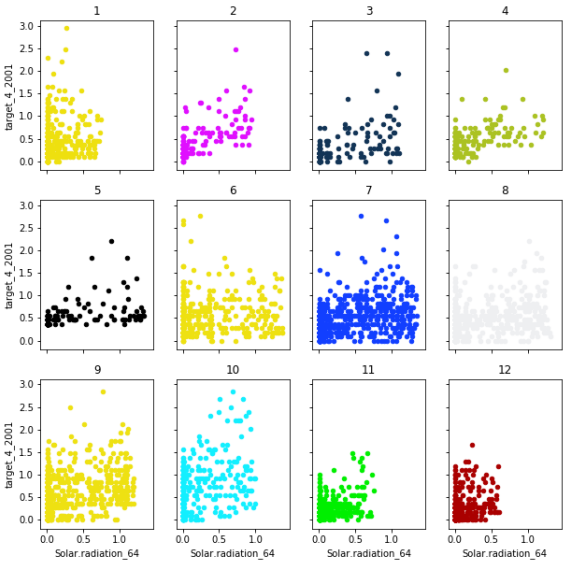


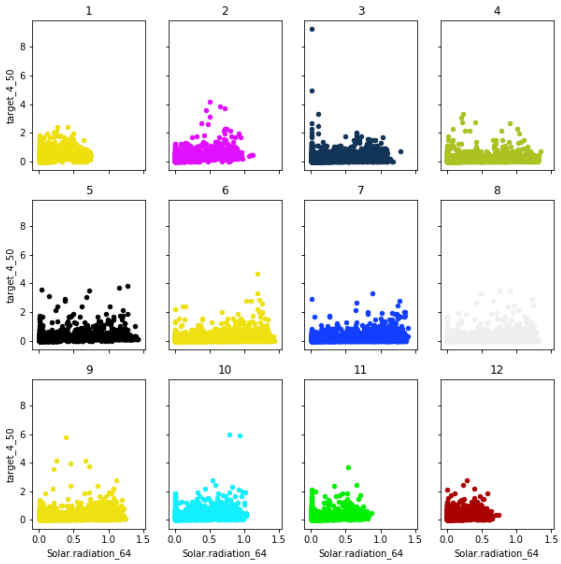


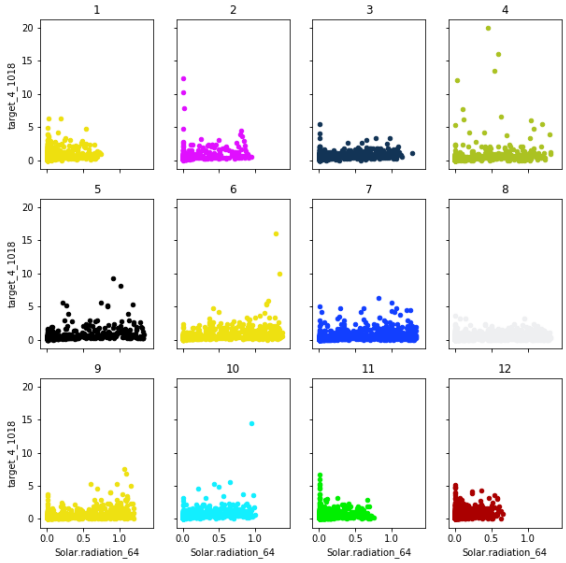


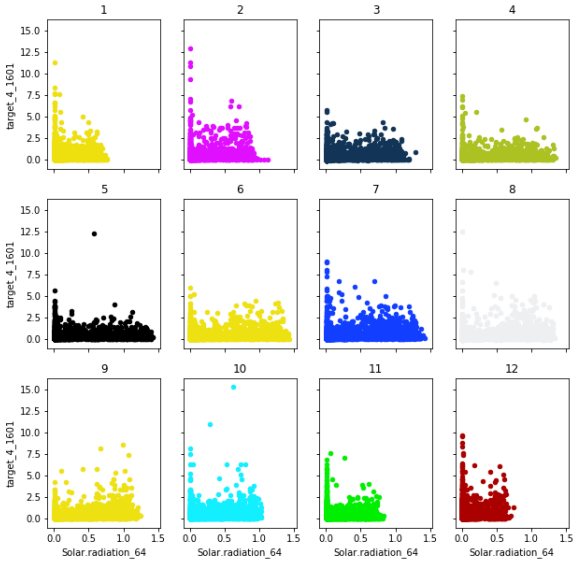


Target 4 vs. Solar Radiation 1

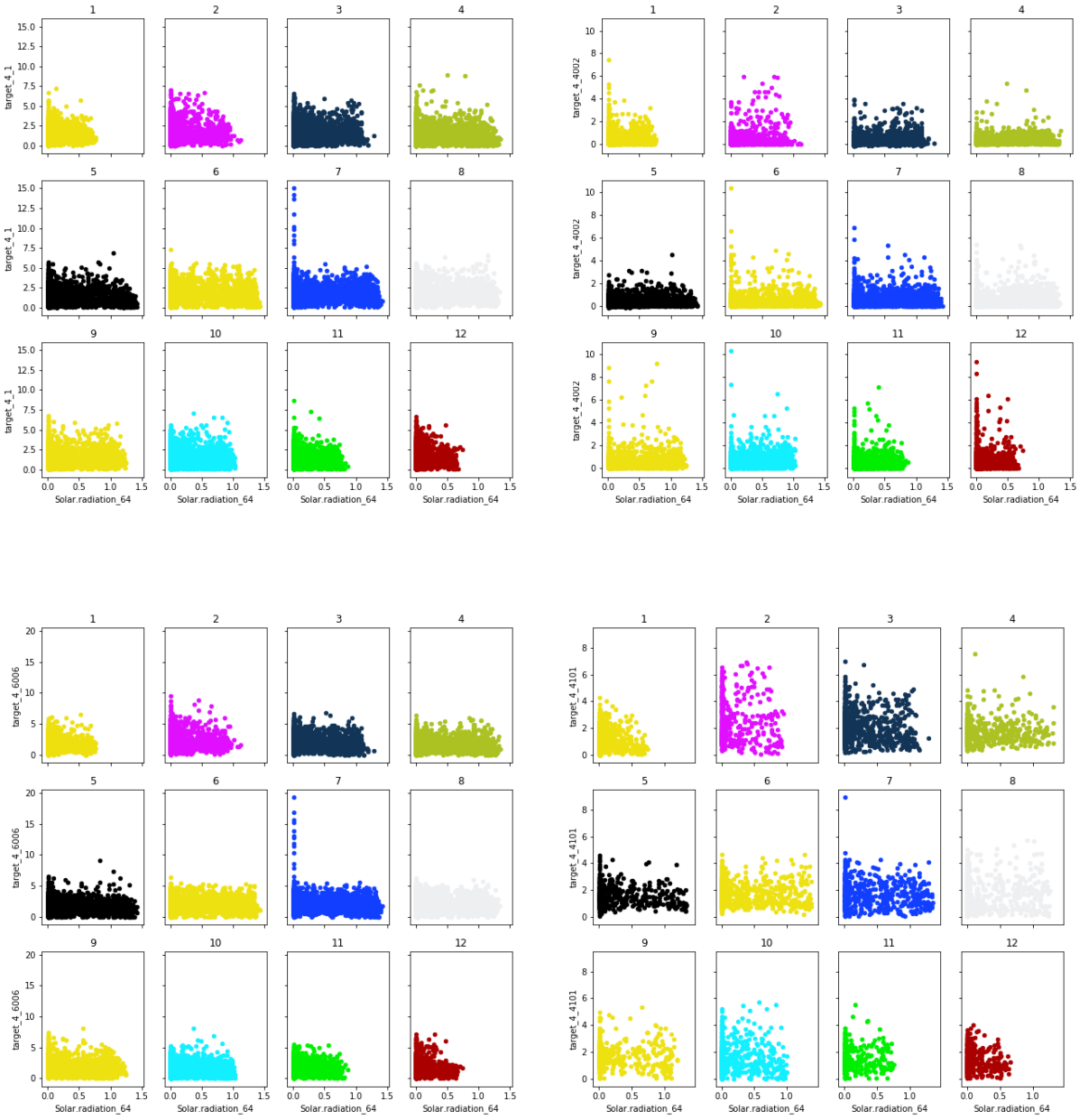




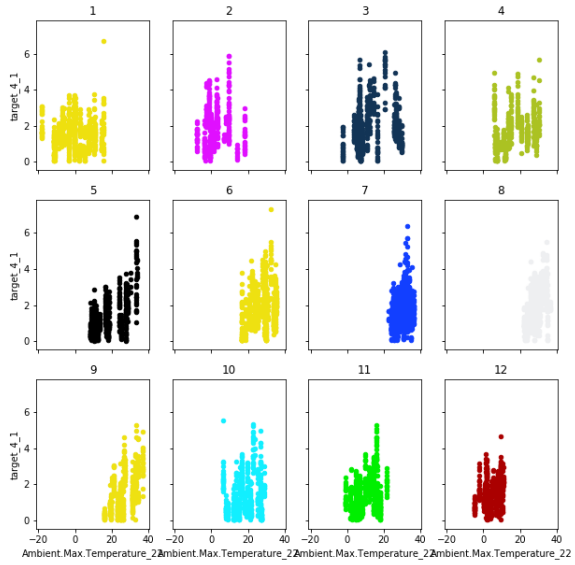


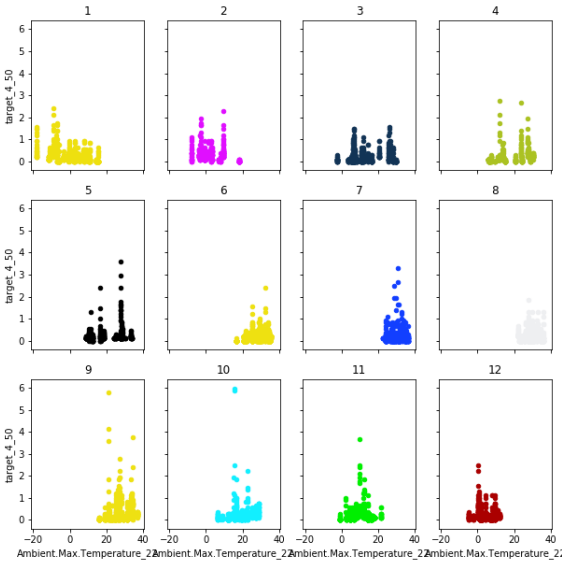


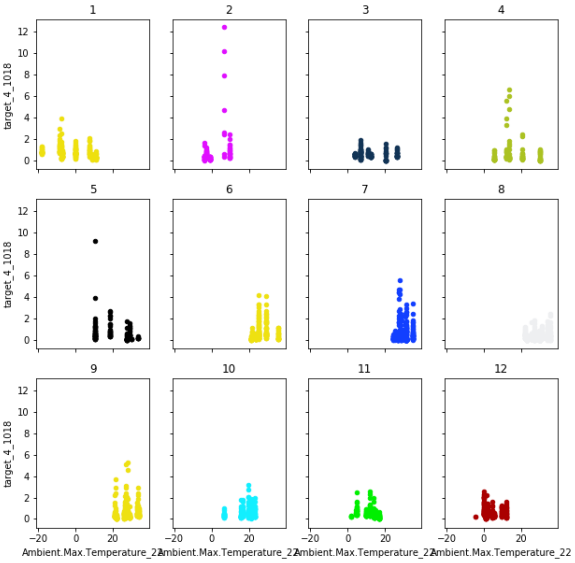
Target 4 vs. Solar Radiation 2

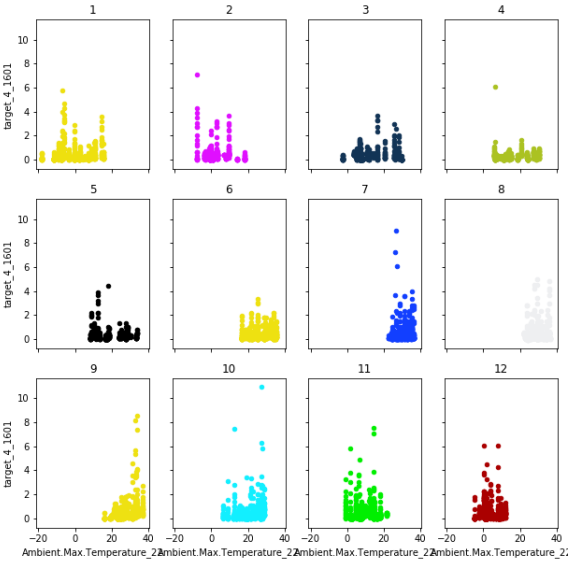


Target 4 vs. temperature 1

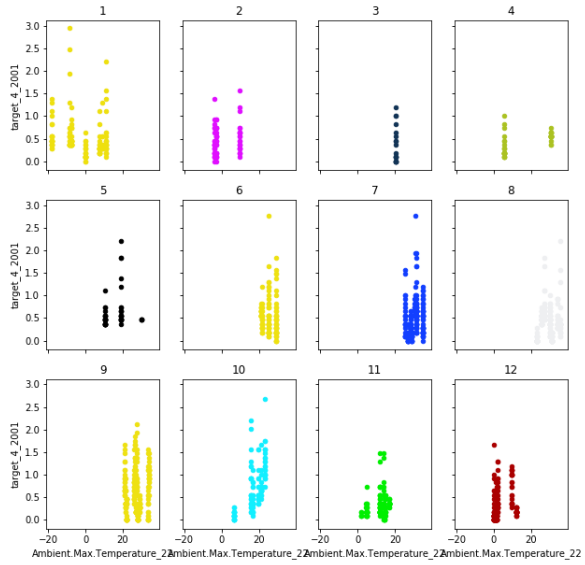


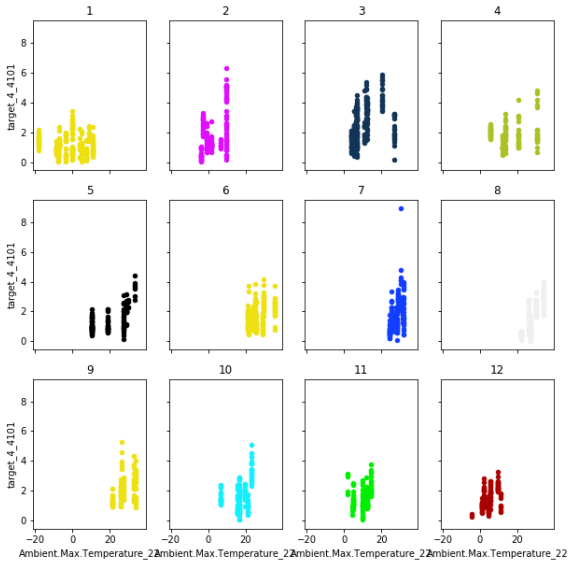


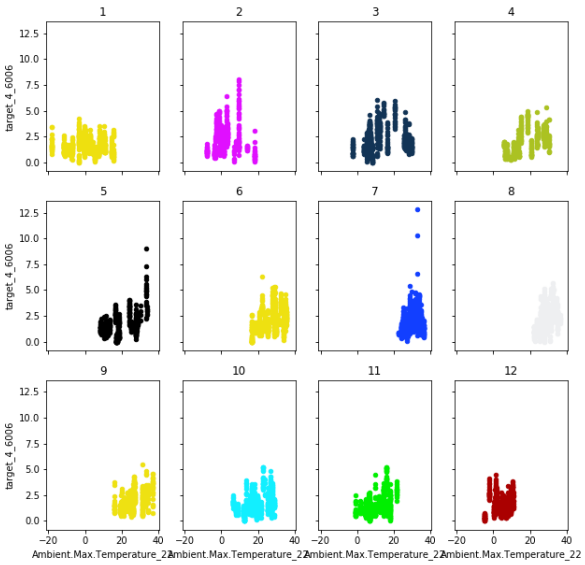




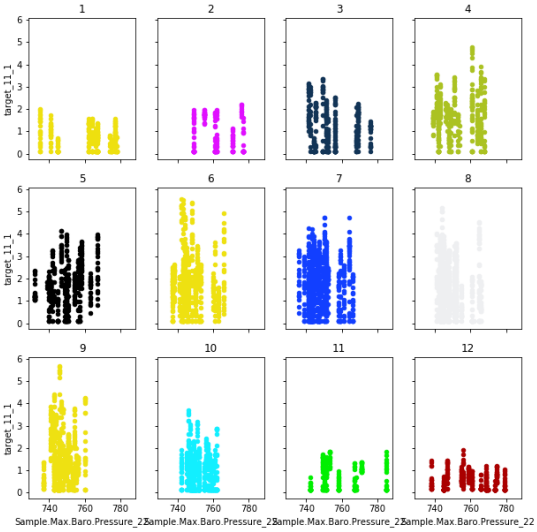
Target 4 vs. temperature 2

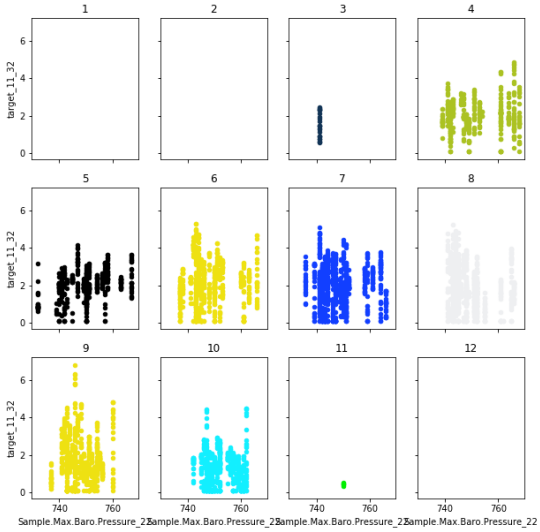


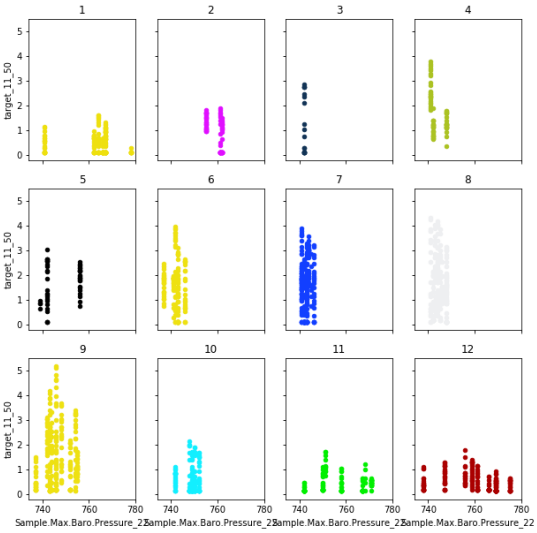


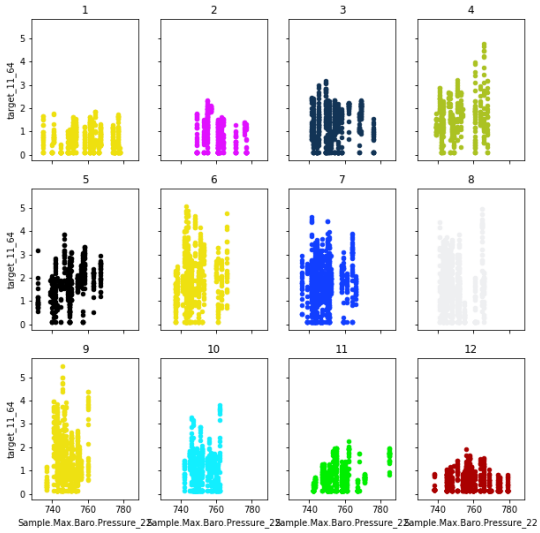


Target 11 vs. Barometric Pressure 1

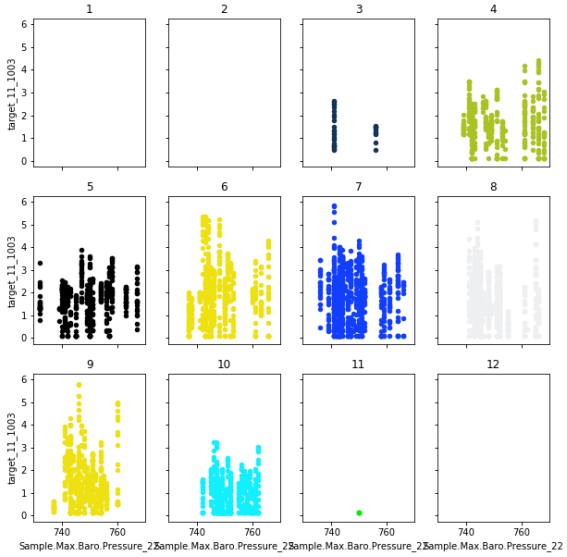


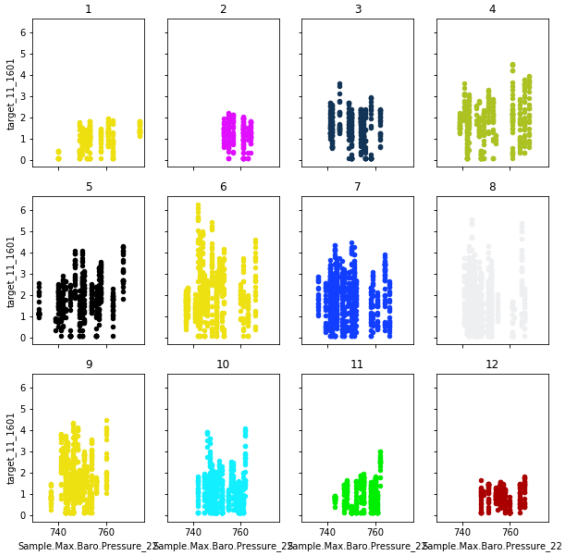


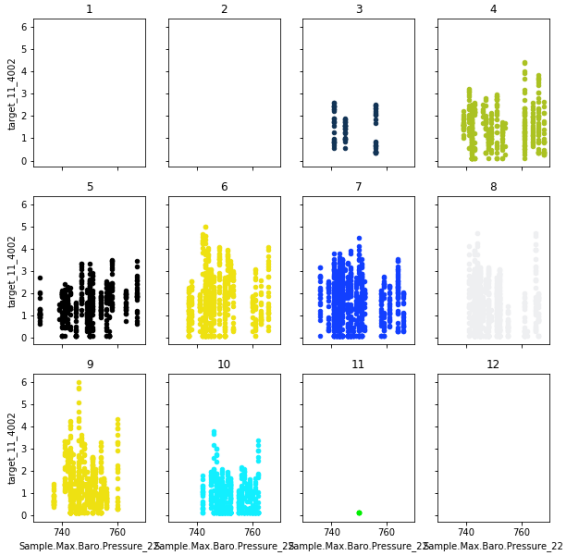


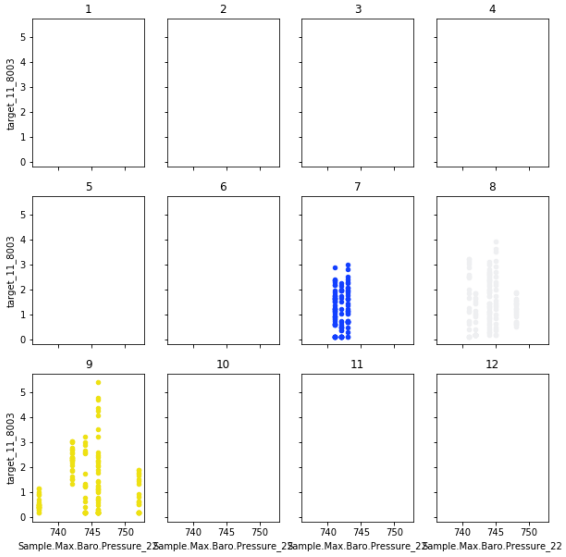


Target 11 vs. Barometric Pressure 2

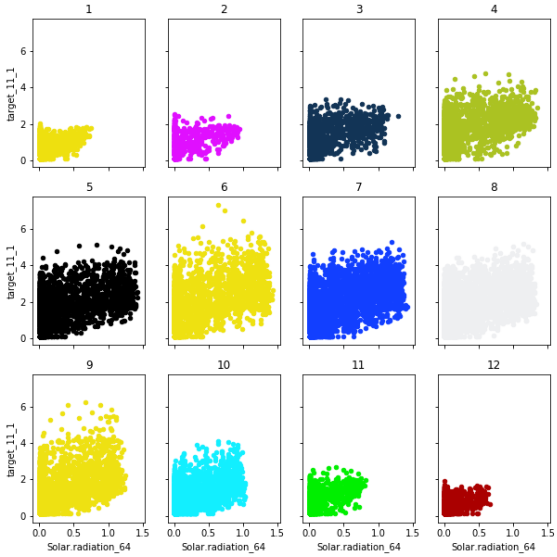


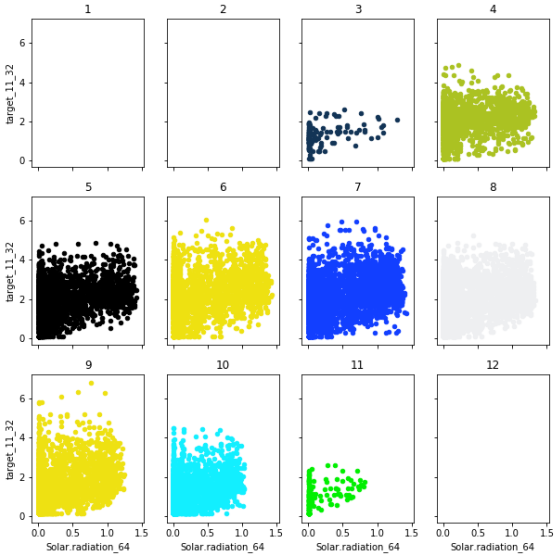


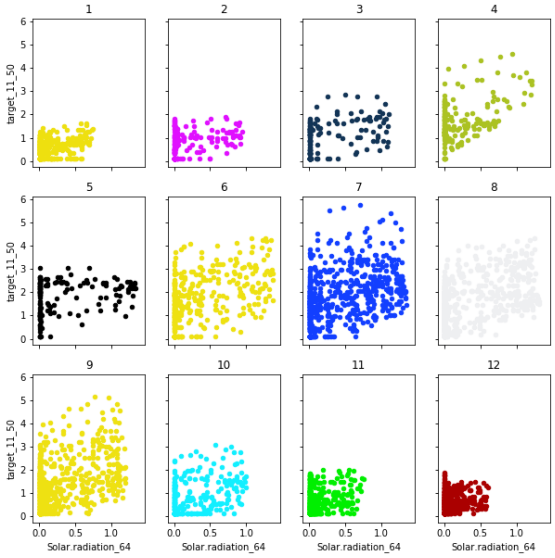


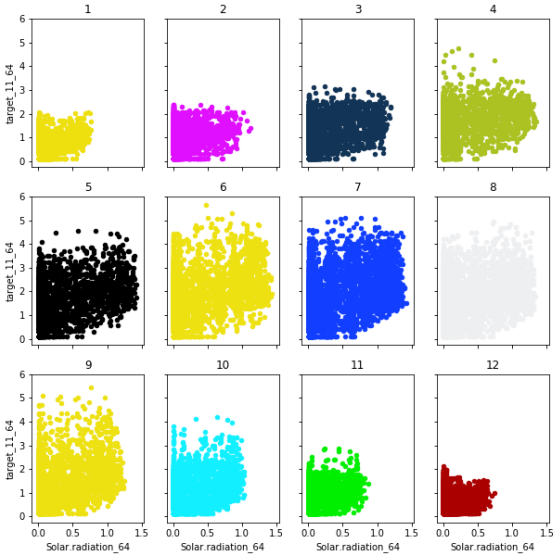


Target 11 vs. Solar Radiation 1

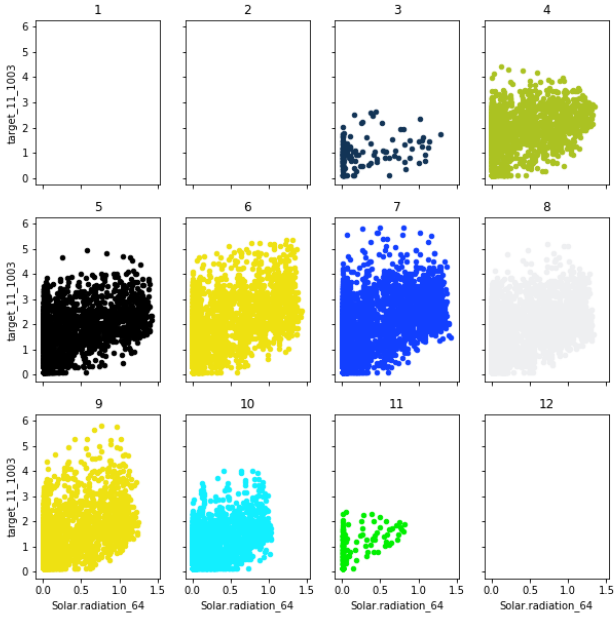


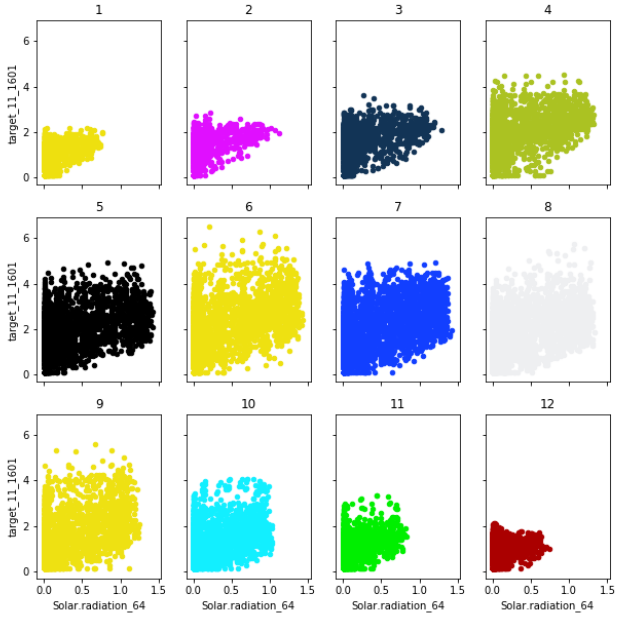


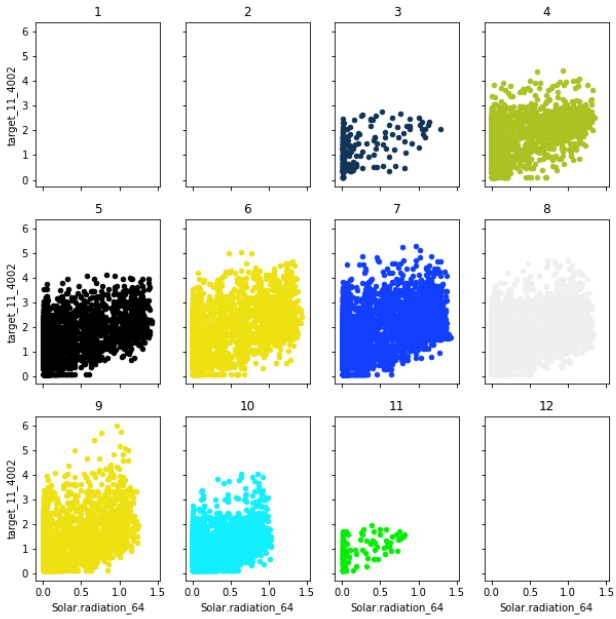


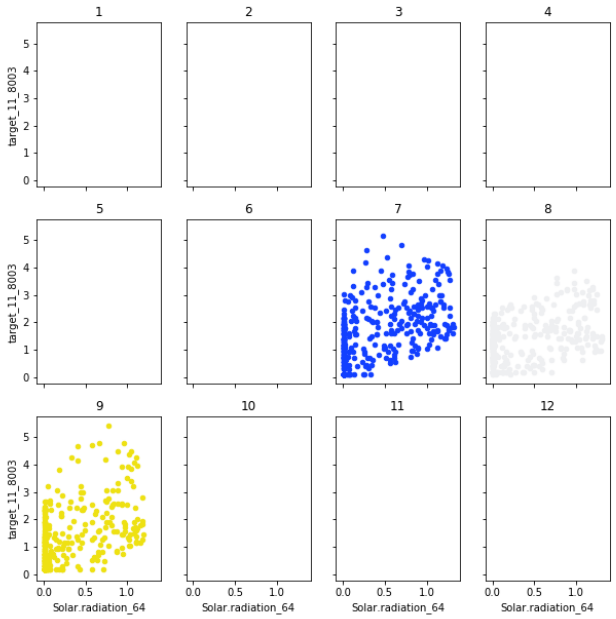


Target 11 vs. Solar Radiation 2

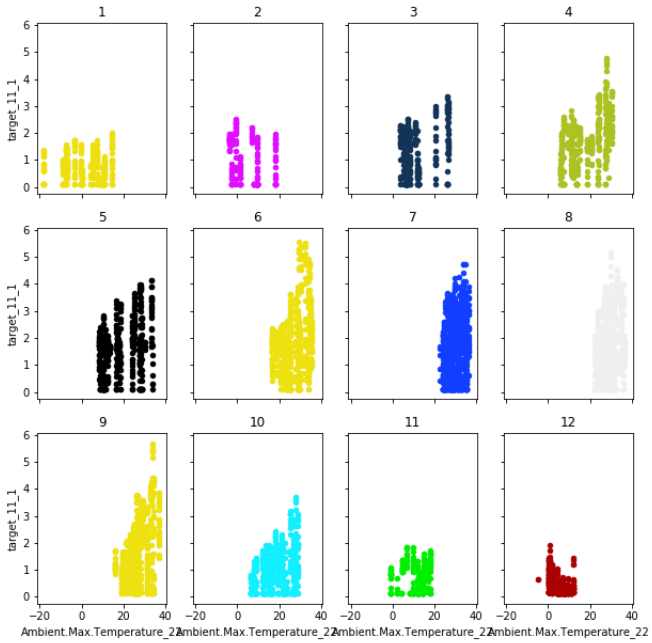


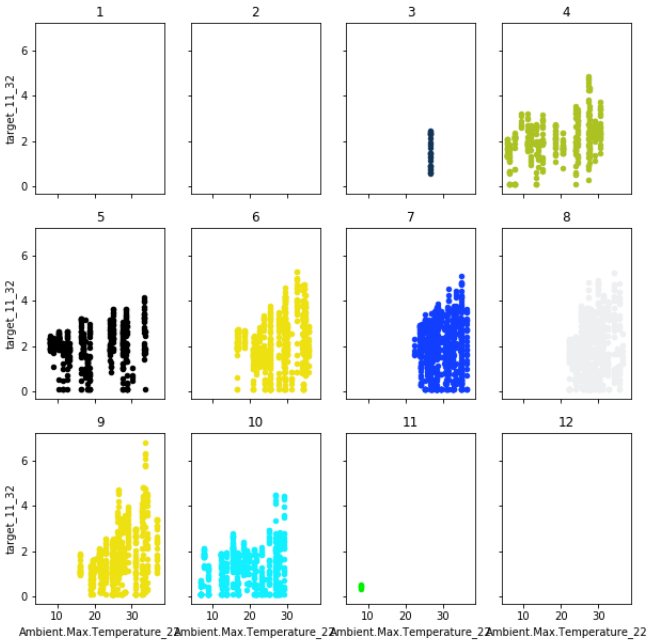


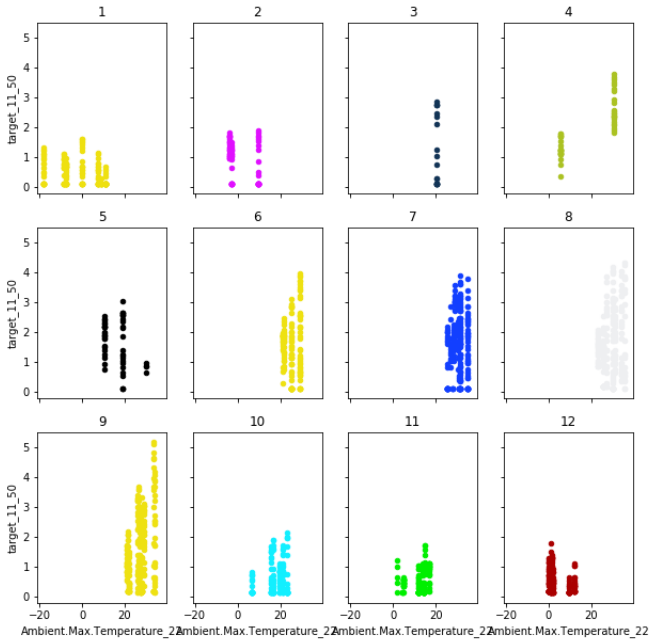


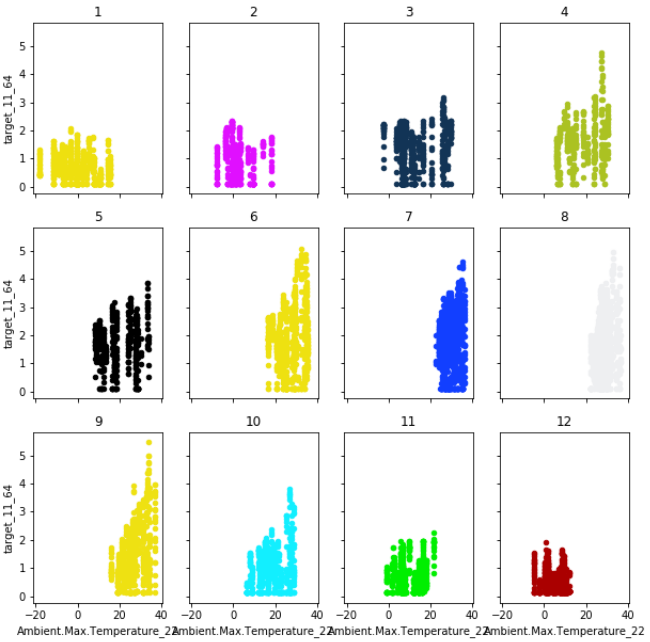


Target 11 vs. Temperature 1

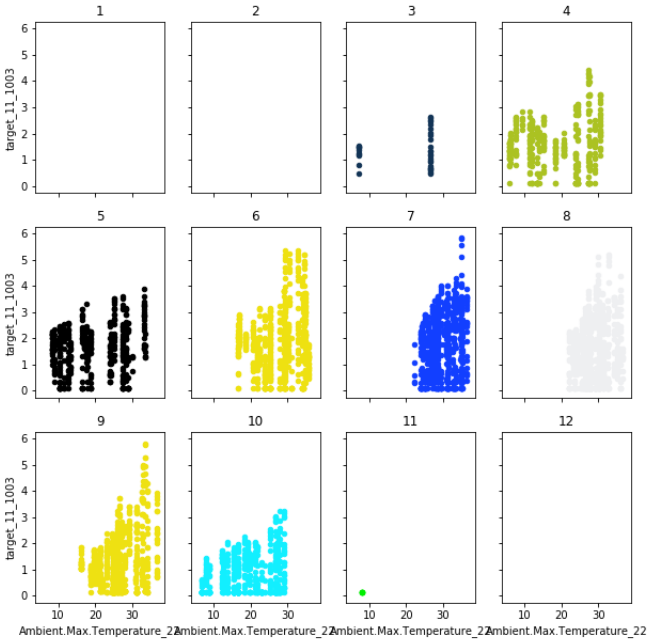


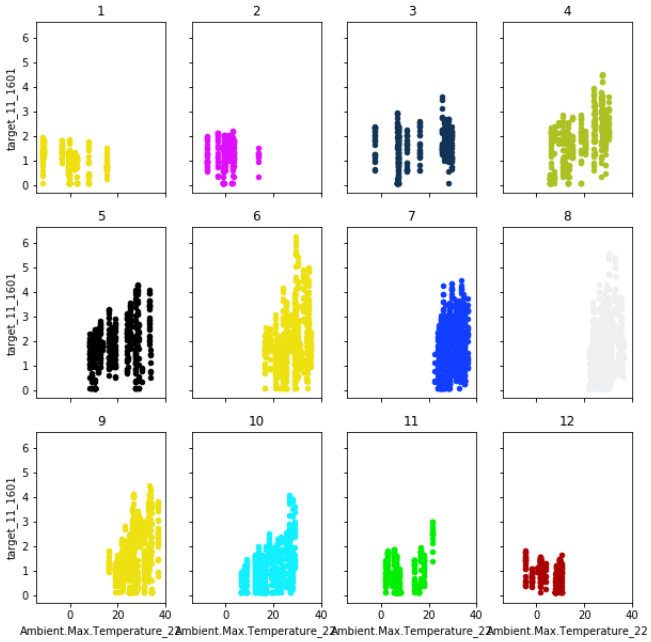


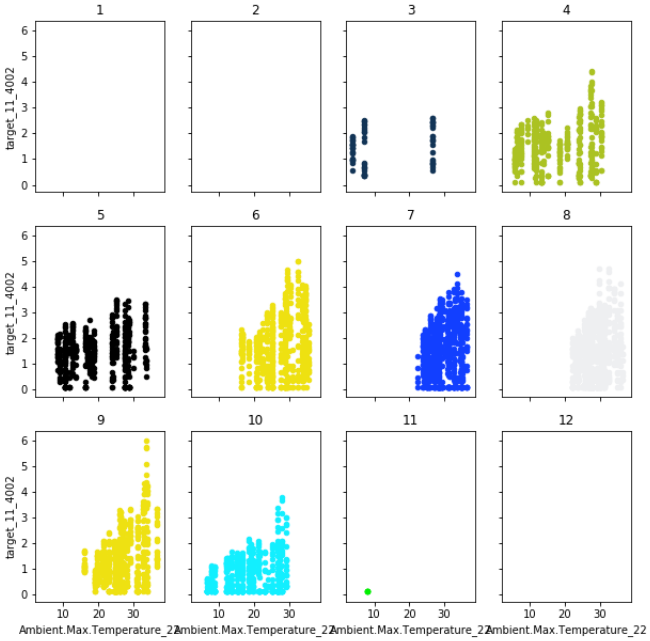


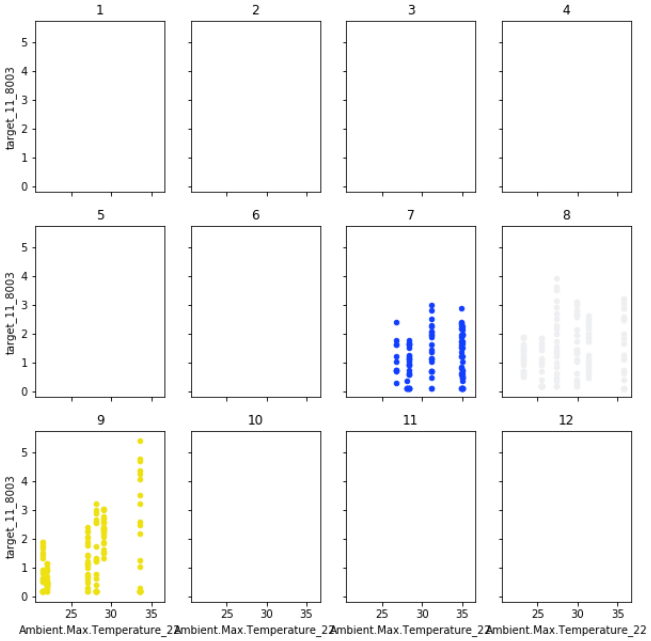


Target 11 vs. Temperature 2

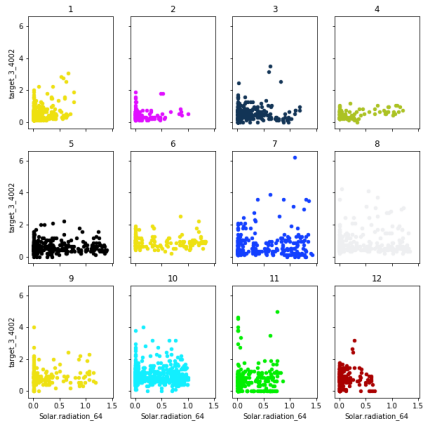


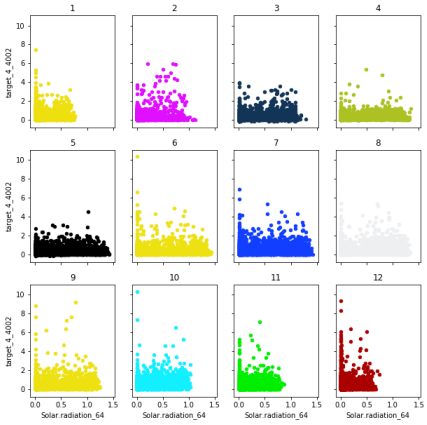


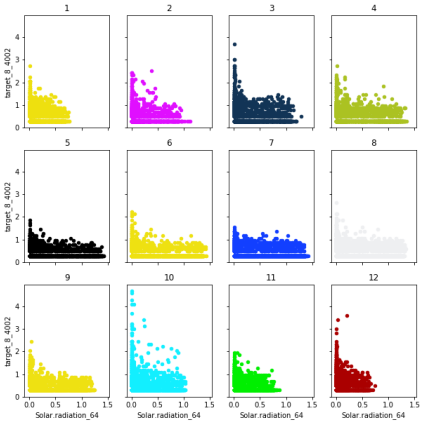


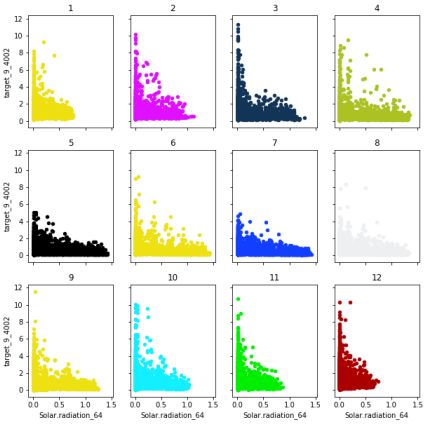


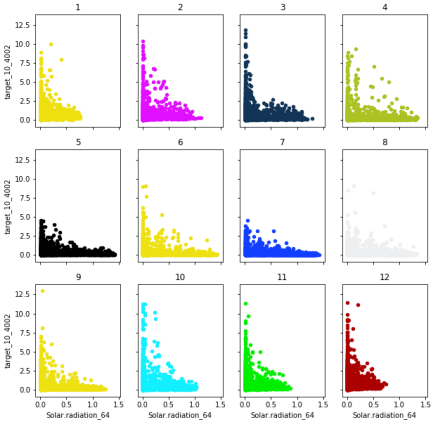
Site 4002 by Solar Radiation 1

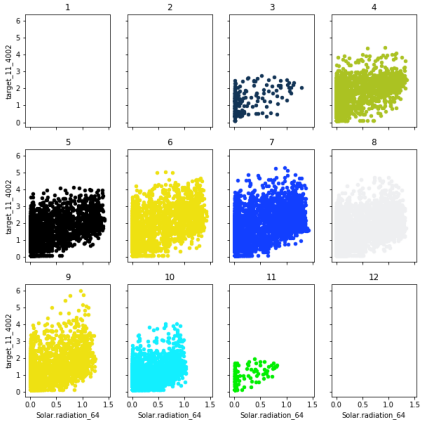


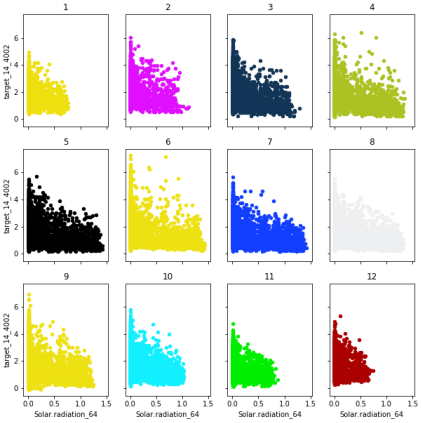




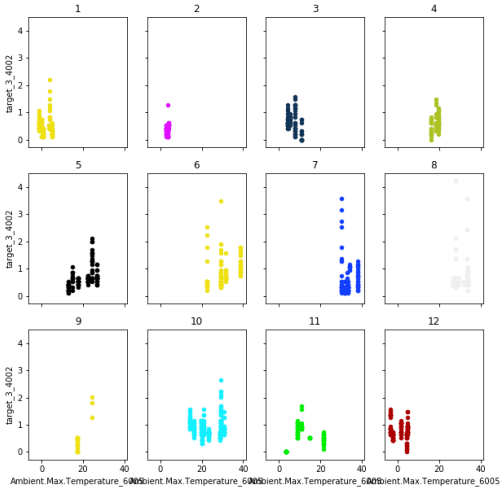


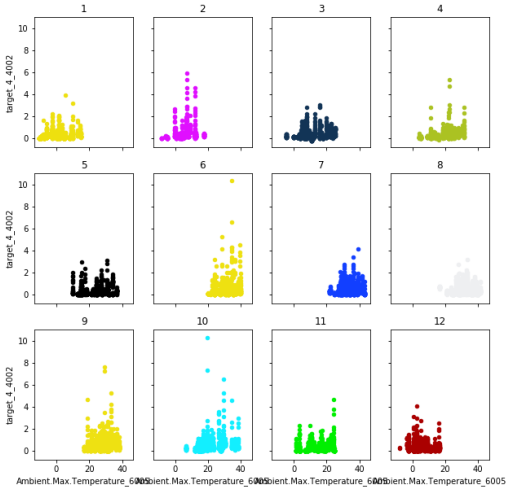


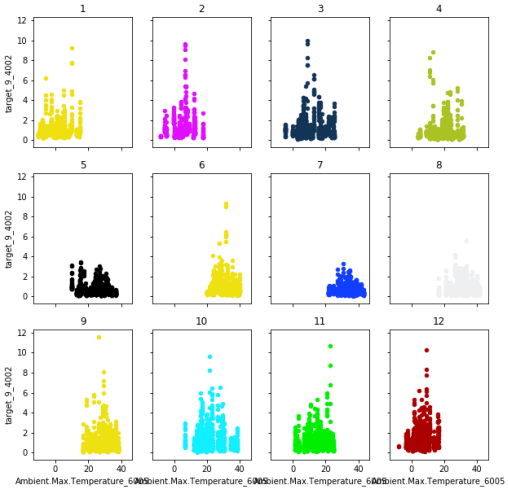


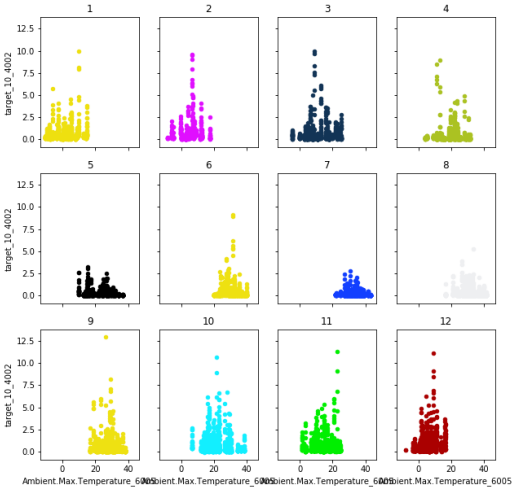


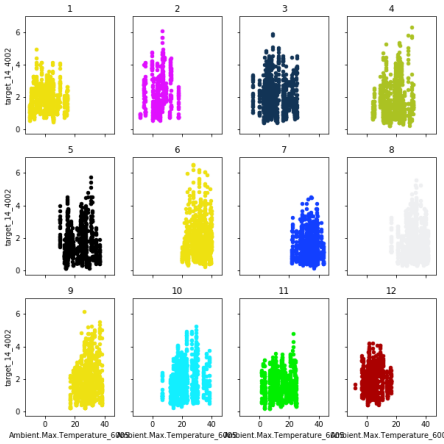
Site 4002 by Temperature 1



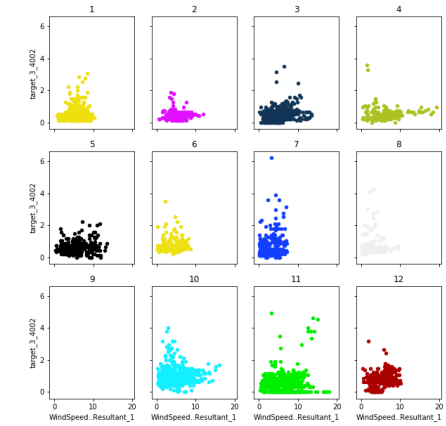


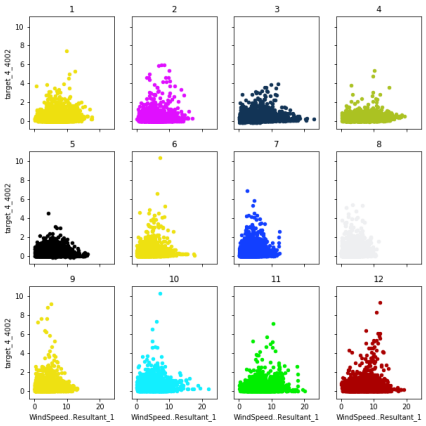


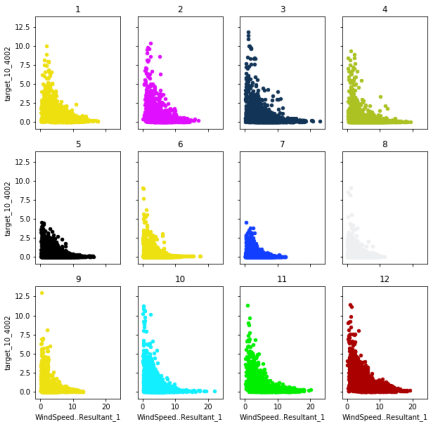


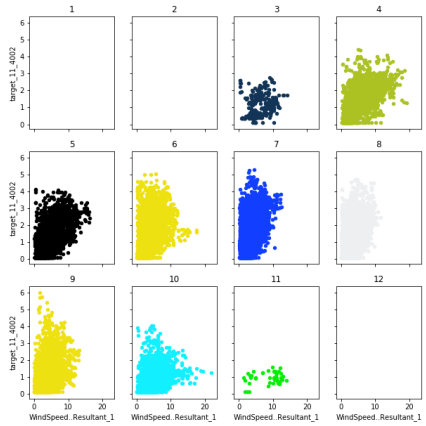


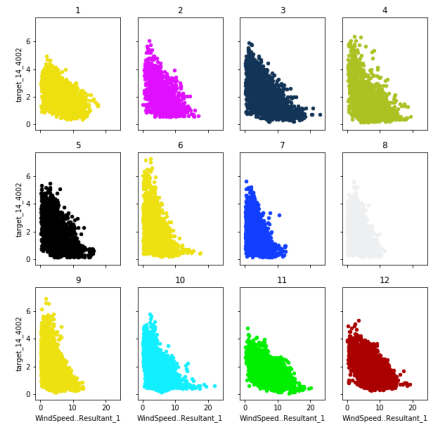
Site 4002 by Windspeed

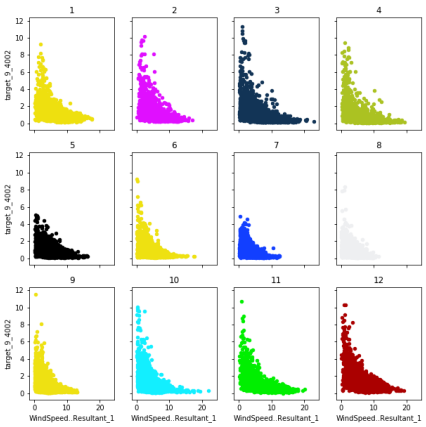




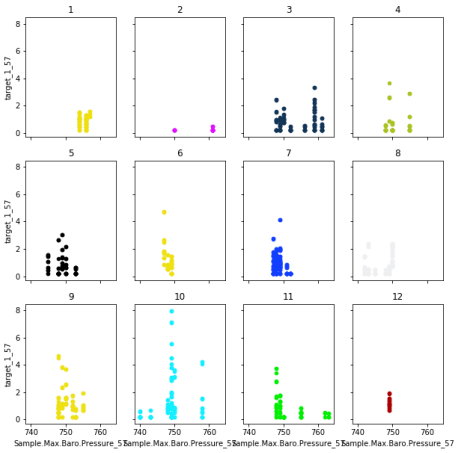


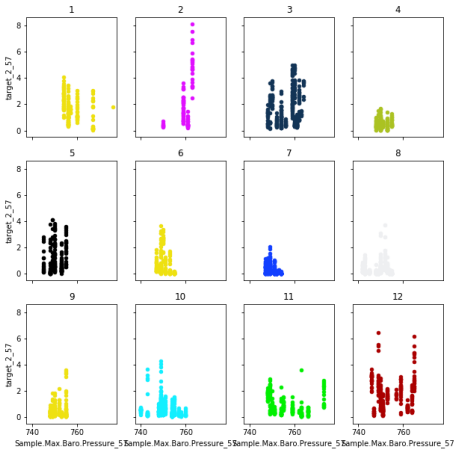


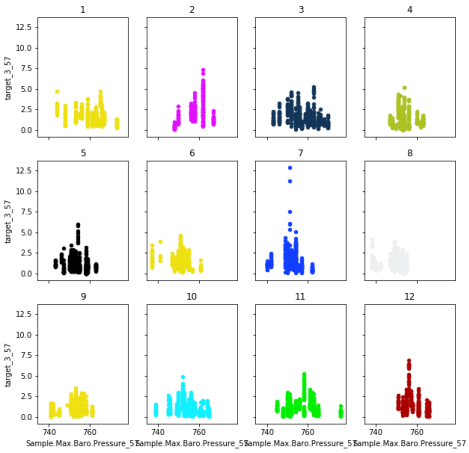


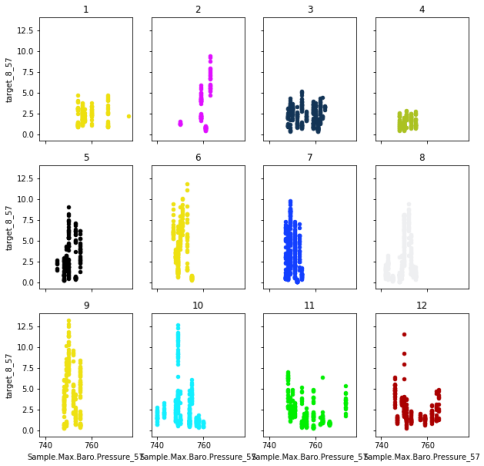


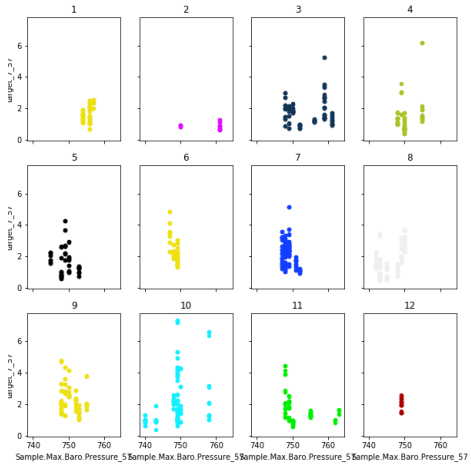
Site 57 by Barometric Pressure

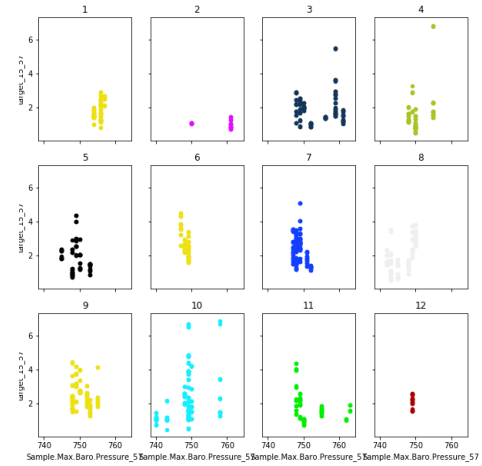


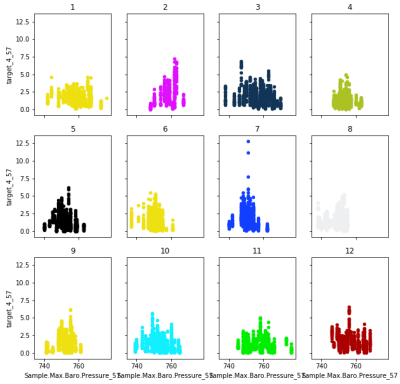




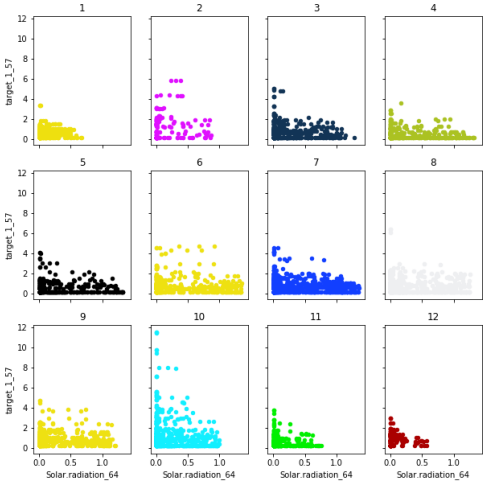


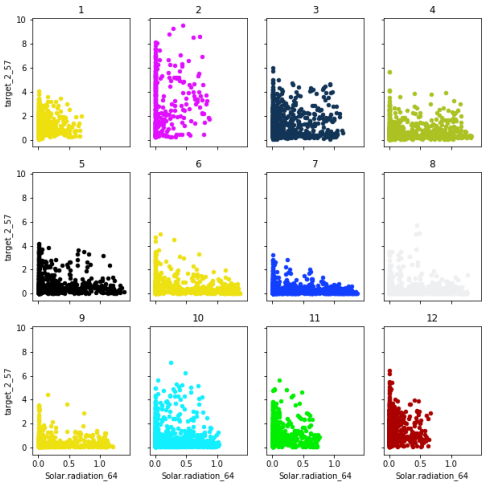


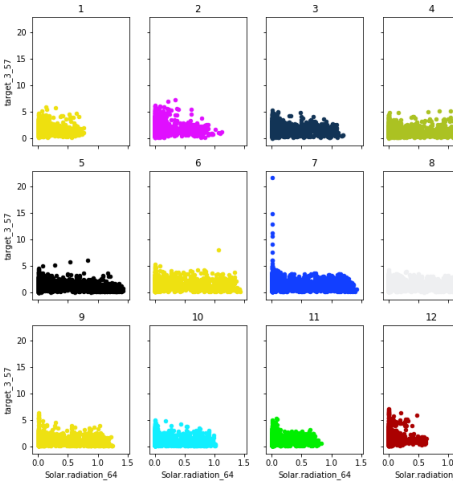


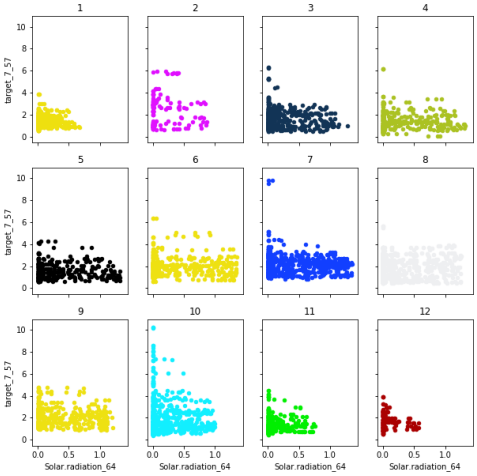


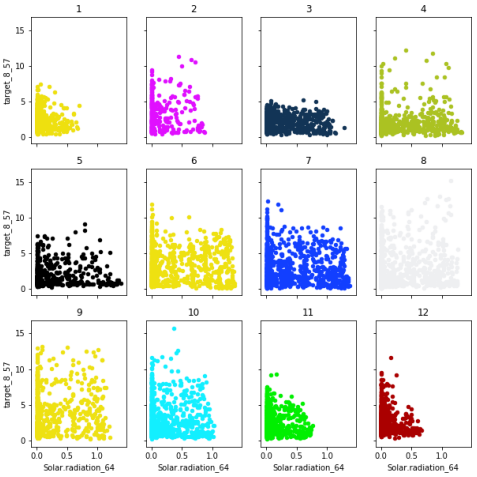
Site 57 by Solar Radiation

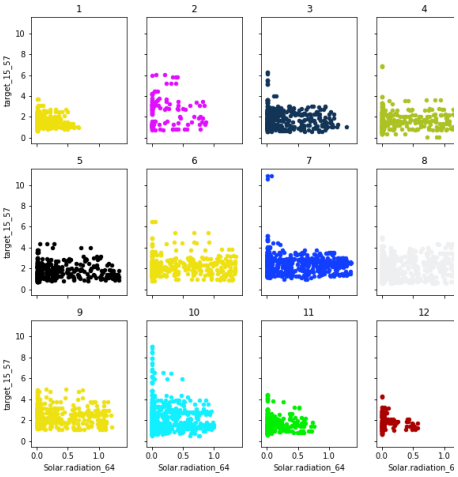


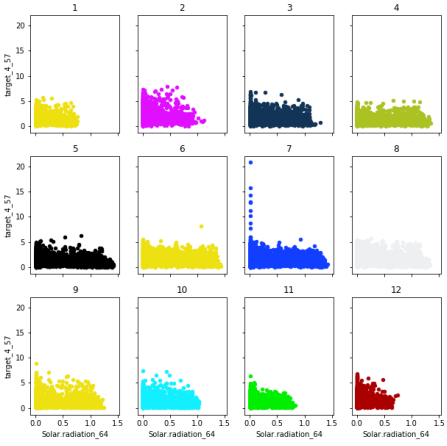




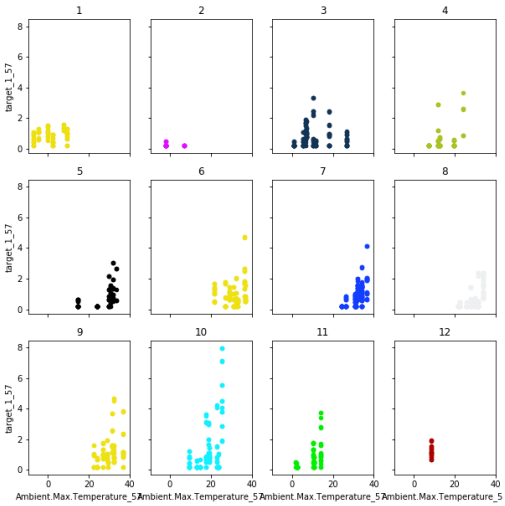


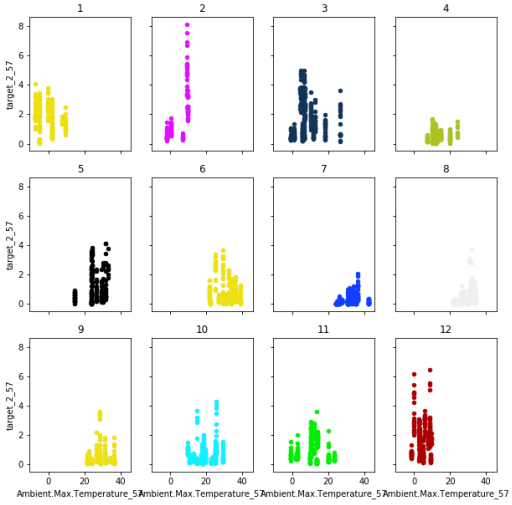


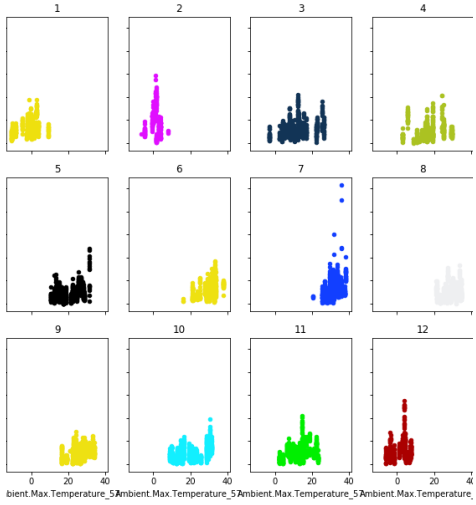


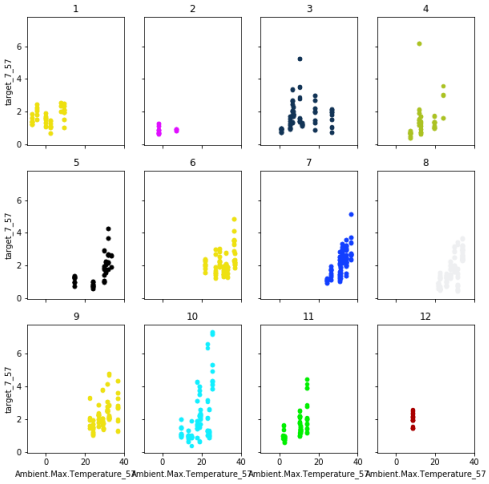


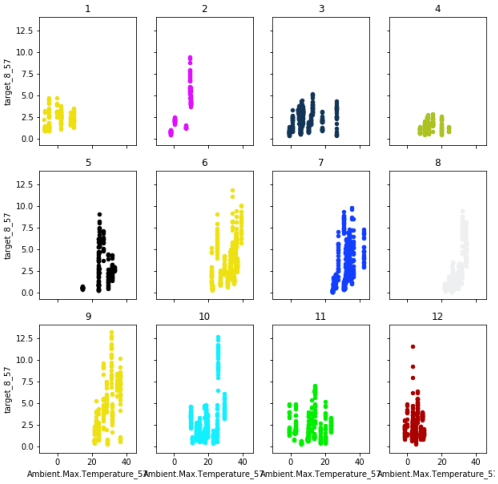
Site 57 by temperature

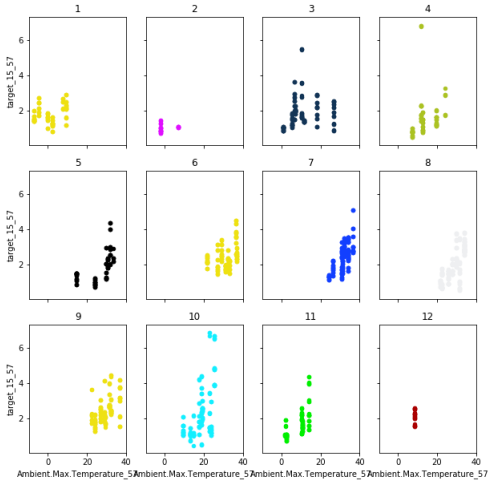


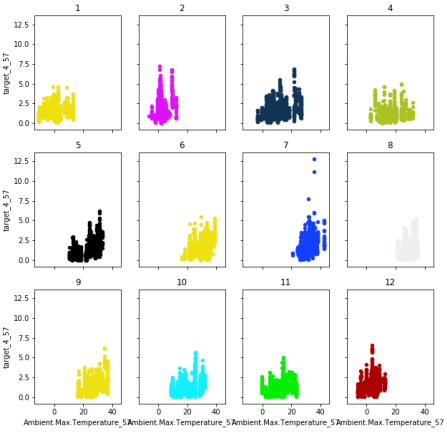




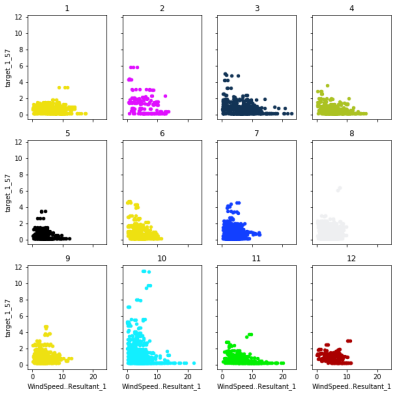


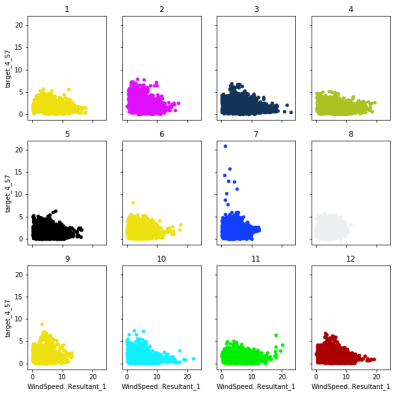


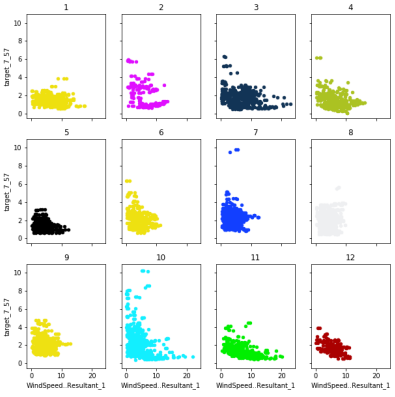


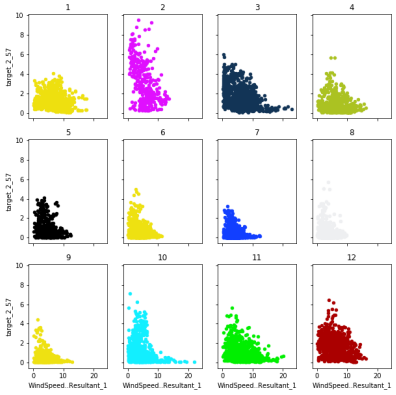


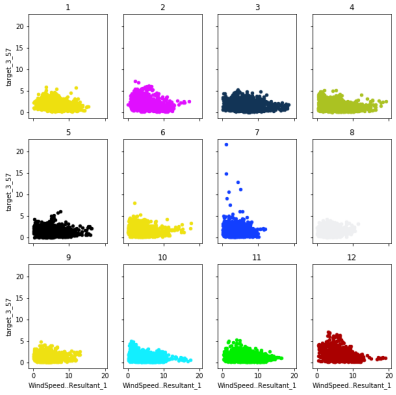
Site 57 by Windspeed

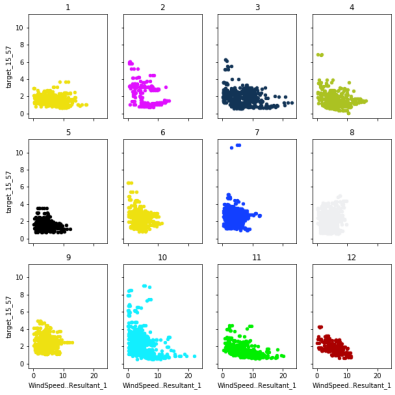


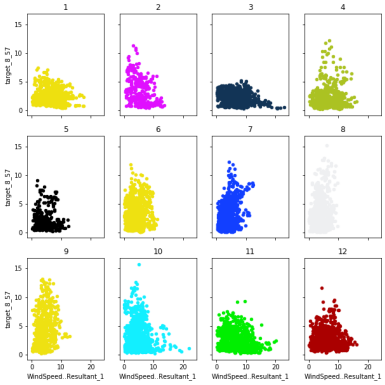












EDA – Targets vs. Targets