PROJECT 5

중합공정라인 데이터 분석 발표

feat. 교차 상관 관계 분석

3팀: yoyoyoyoyo

정은서 | 이예솔 | 장일준
김연진 | 이다경

목차 LIST

01 분석 주제 소개

- **분석 주제**

색조 L치 변수의 이상에 영향을 준 변수와 시기 알아보기

*변수명: LAB_8CHIP_L

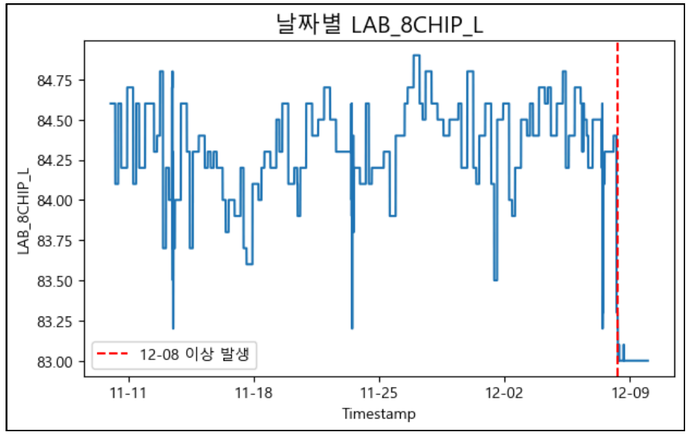
- **수집 데이터**

중합 공정라인(CPS-8)에서 수집한 데이터

- **데이터 수집기간**

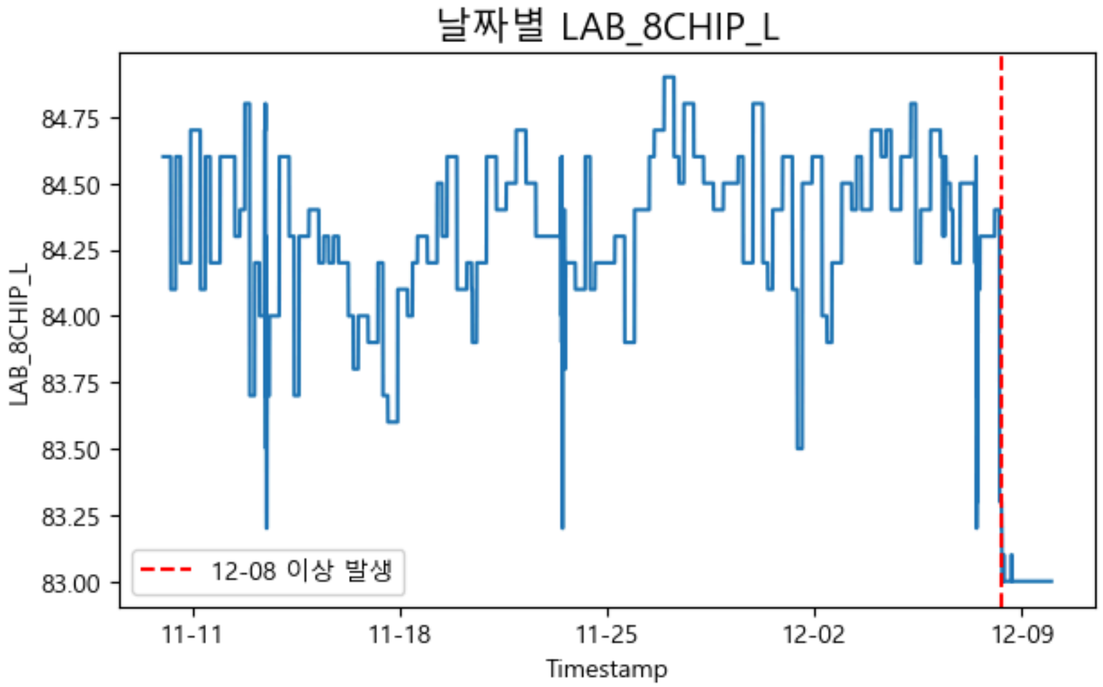
2018년 11월 10일 00시 00분 ~

2018년 12월 10일 00시 00분 (1분 단위)

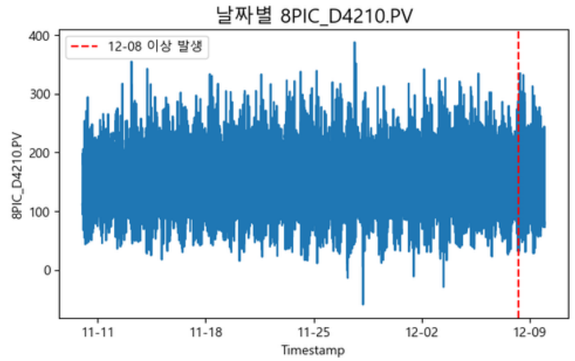


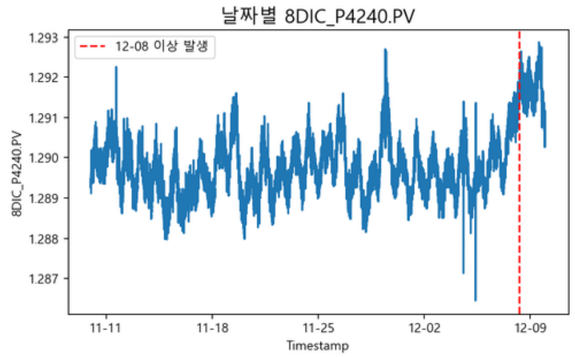
01 분석 주제 소개

타깃 변수



설명 변수





데이터분석 진행과정

STEP 01

데이터 소개 전처리

EDA & 전처리

- 중복 행, 열 제거
- 단독 날짜 행 11/9 제거
- 단일 값 갖는 열 제거

STEP 02

교차분석

CCF 활용

설명변수와 타깃과의 상관 관계 확인

- 이산형 (LAB)
- 연속형

STEP 03

회귀분석

단순 선형 회귀 모델

CCF로 도출된 주요 변수와 타깃과의 인과 관계를 통계적으로 검정(p-value)

STEP 04

결론도출

원인변수, 원인시점

모니터링 필요 변수 확인

02 데이터 소개 및 전처리

02 데이터 소개 및 전처리

B1

중합 CPS-8 DB B1.xlsx
44641행, 199열
날짜 변수 1개, 그외 모두 수치 변수
결측치 없음

B2

중합 CPS-8 DB B2.xlsx
43201행, 213열
날짜 변수 1개, 그 외 모두 수치 변수
결측치 없음

```
df_B1.head()
```

	Timestamp	PIC_D4210.PV - Average	8DIC_P4240.PV - Average	8XIC_P4260.PV - Average	8XIC_P4230.PV - Average	8SIC_M4220.PV - Average	8FIC_P4260B.PV - Average
0	2018-11-10 00:00:00	111.946831	1.289244	2.554061	2.211044	15.034183	3422.315023
1	2018-11-10 00:01:00	112.259820	1.289276	2.553664	2.210712	15.026483	3426.222453
2	2018-11-10 00:02:00	109.442167	1.289382	2.559363	2.210777	14.930664	3410.658244
3	2018-11-10 00:03:00	106.624023	1.289399	2.550211	2.209736	14.947618	3409.020711
4	2018-11-10 00:04:00	103.962240	1.289293	2.551153	2.210752	14.946071	3395.153727

```
df_B2.head()
```

	Timestamp	TI_P4360A.PV - Average	8TI_P4360B.PV - Average	8II_P4360A.PV - Average	8II_P4360B.PV - Average	8PI_P4360A.PV - Average	8PI_P4360B.PV - Average
0	2018-11-10 00:00:00	255.299435	282.387547	78.845353	82.753790	37.561569	41.452562
1	2018-11-10 00:01:00	255.276543	282.371912	78.845666	82.747861	37.544696	41.450687
2	2018-11-10 00:02:00	255.287989	282.376003	78.833184	82.765958	37.560319	41.480685
3	2018-11-10 00:03:00	255.294267	282.368925	78.813214	82.811202	37.565320	41.463187
4	2018-11-10 00:04:00	255.321214	282.384198	78.851594	82.907306	37.577194	41.519432

=> ‘Timestamp’를 기준으로 열 병합하기 위해 작업 필요

02 데이터 소개 및 전처리

- 두 데이터셋 행 맞추기

B2에 없는 날짜
“2018-11-09” 제거

B1

	Timestamp	8PIC_D4210.PV - Average
43201	2018-11-09 00:01:00	179.896591
43202	2018-11-09 00:02:00	208.392253
43203	2018-11-09 00:03:00	208.392253
43204	2018-11-09 00:04:00	209.957954
43205	2018-11-09 00:05:00	212.305936
...	...	...
44640	2018-11-10 00:00:00	111.946831
1	2018-11-10 00:01:00	112.259820
2	2018-11-10 00:02:00	109.442167
3	2018-11-10 00:03:00	106.624023
4	2018-11-10 00:04:00	103.962240

B2

	Timestamp	8TI_P4360A.PV - Average
0	2018-11-10 00:00:00	255.299435
1	2018-11-10 00:01:00	255.276543
2	2018-11-10 00:02:00	255.287989
3	2018-11-10 00:03:00	255.294267
4	2018-11-10 00:04:00	255.321214
...	...	...

02 데이터 소개 및 전처리

- **중복 행 제거**

  B1 1행 제거

	Timestamp	8PIC_D4210.PV - Average	8DIC_P4240.PV - Average	8XIC_P4260.PV - Average	8XIC_P4230.PV - Average	8SIC_M4220.PV - Average	8FIC_P4260B.PV - Average
0	2018-11-10	111.946831	1.289244	2.554061	2.211044	15.034183	3422.315023
44640	2018-11-10	111.946831	1.289244	2.554061	2.211044	15.034183	3422.315023

- **중복 열 제거**

  B1 4열 제거

	8FIC_P4260A.PV - Average	8FIC_P4260B.PV - Average	8FIC_P4230A.PV - Average	8FIC_P4230B.PV - Average	8TIC_E9340.PV - Average	8TIC_E9340.PV - Average.1	8FIC_F4270A.PV - Average	8FIC_F4270B.PV - Average
0	3422.315023	3422.315023	4138.207804	4138.207804	272.640216	272.640216	11.101759	11.101759
1	3426.222453	3426.222453	4141.679443	4141.679443	272.645431	272.645431	11.096759	11.096759
2	3410.658244	3410.658244	4118.675700	4118.675700	272.656982	272.656982	11.105793	11.105793
3	3409.020711	3409.020711	4119.256226	4119.256226	272.673673	272.673673	11.110695	11.110695
4	3395.153727	3395.153727	4112.225301	4112.225301	272.664787	272.664787	11.109785	11.109785

- **고유값 1개인 변수 제거**

  B1 3열, B2 2열 제거

8PI_P4251A.PV - Average	8FI_D8502.PV - Average	8FI_F4270B.PV - Average
-0.25	123	0
-0.25	123	0
-0.25	123	0
-0.25	123	0

LAB_8CHIP_TIO2 - Average	LAB_8CHIP_HEAT - Average
0.271	2.5
0.271	2.5
0.271	2.5
0.271	2.5

02 데이터 소개 및 전처리

- B1, B2 열 병합 => B (행 43201, 열 396)

	Timestamp	8PIC_D4210.PV - Average	8DIC_P4240.PV - Average	8XIC_P4260.PV - Average	8XIC_P4230.PV - Average	8SIC_M4220.PV - Average
0	2018-11-10 00:00:00	111.946831	1.289244	2.554061	2.211044	15.034183
1	2018-11-10 00:01:00	112.259820	1.289276	2.553664	2.210712	15.026483
2	2018-11-10 00:02:00	109.442167	1.289382	2.559363	2.210777	14.930664

(행 43201, 열 195)

+

	Timestamp	TI_P4360A.PV - Average	8TI_P4360B.PV - Average	8II_P4360A.PV - Average	8II_P4360B.PV - Average	8PI_P4360A.PV - Average
0	2018-11-10 00:00:00	255.299435	282.387547	78.845353	82.753790	37.561569
1	2018-11-10 00:01:00	255.276543	282.371912	78.845666	82.747861	37.544696
2	2018-11-10 00:02:00	255.287989	282.376003	78.833184	82.765958	37.560319

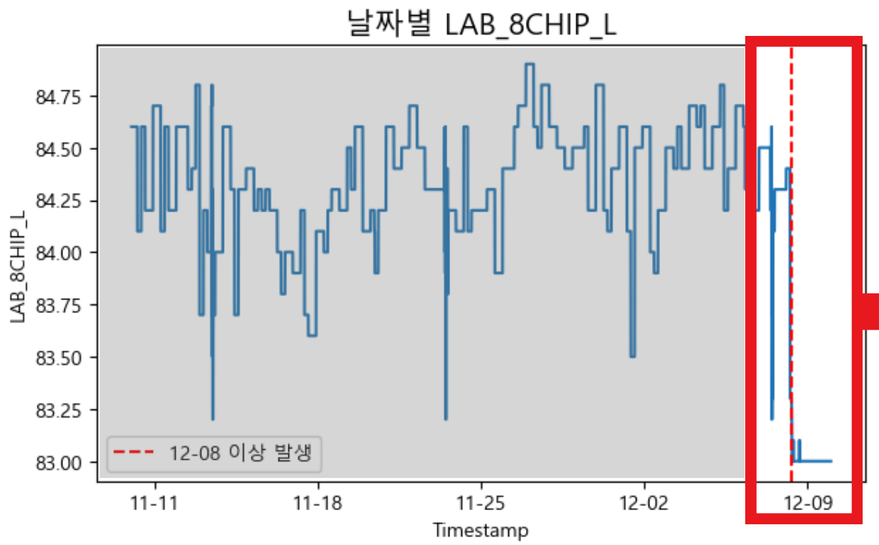
(행 43201, 열 213)

- 변수명 바꾸기 => “- Average” 제거

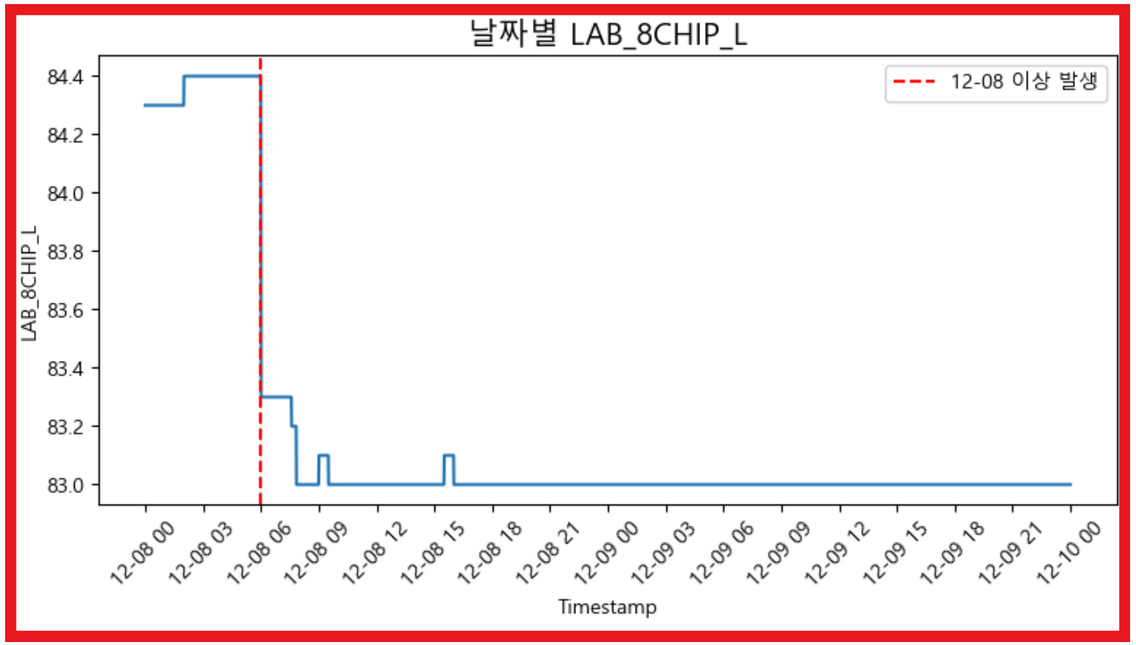
	Timestamp	8PIC_D4210.PV - Average	8DIC_P4240.PV - Average	8XIC_P4260.PV - Average	8XIC_P4230.PV - Average	8SIC_M4220.PV - Average
0	2018-11-10 00:00:00	111.946831	1.289244	2.554061	2.211044	15.034183
1	2018-11-10 00:01:00	112.259820	1.289276	2.553664	2.210712	15.026483
2	2018-11-10 00:02:00	109.442167	1.289382	2.559363	2.210777	14.930664

	Timestamp	8PIC_D4210.PV	8DIC_P4240.PV	8XIC_P4260.PV	8XIC_P4230.PV	8SIC_M4220.PV
0	2018-11-10 00:00:00	111.946831	1.289244	2.554061	2.211044	15.034183
1	2018-11-10 00:01:00	112.259820	1.289276	2.553664	2.210712	15.026483
2	2018-11-10 00:02:00	109.442167	1.289382	2.559363	2.210777	14.930664

02 데이터 소개 및 전처리



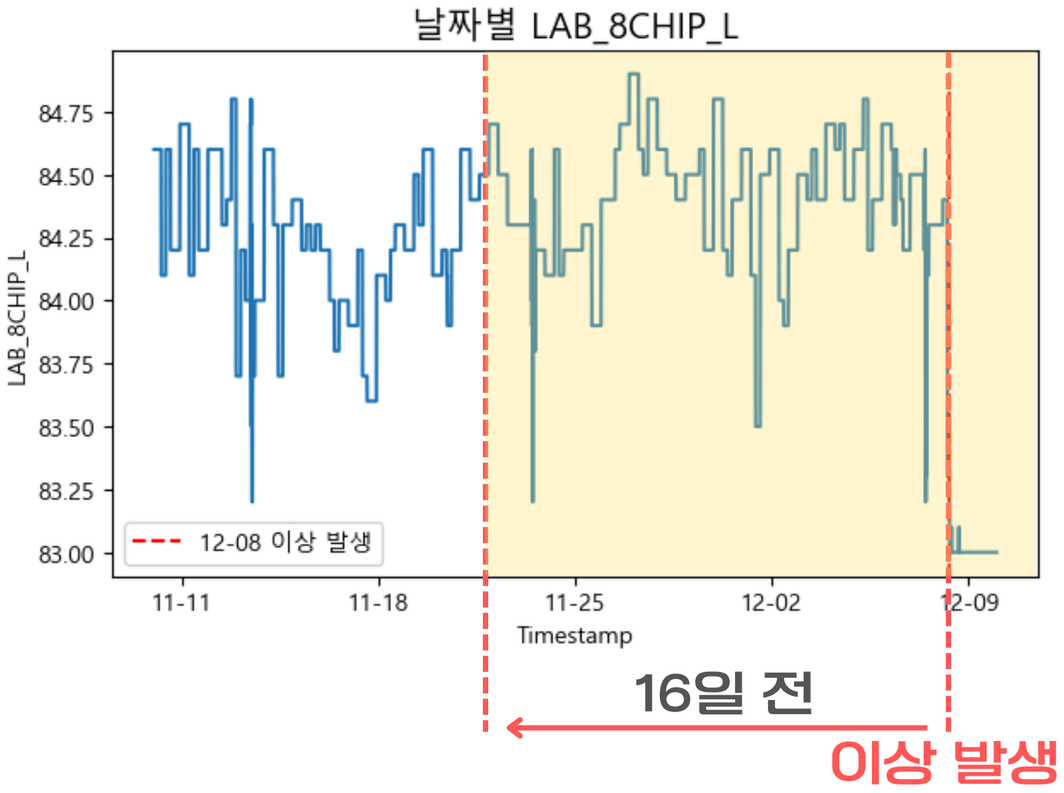
전체기간 중 타겟변수 시계열 그래프



12월 8일 부터 타겟변수 시계열 그래프

이상현상에 상관이 높은 변수 찾고자 함

02 데이터 소개 및 전처리



분석 데이터 범위 설정

- 공정 1사이클이 16일이라고 가정
- 이상 발생 시점 : 2018-12-08 06:00:00
- 이상 발생 16일 전: 2018-11-22 06:00:00
- 이상 16일 전 이후 데이터를 기준으로 분석 진행
- 이상 발생 이전에 영향을 미친 요인을 찾기 위해 교차 분석으로 변수 선정 시 lag0인 값은 제외

03 교차 상관 분석

03 교차 상관 분석

- 교차 상관 분석 : 두 변수의 **시간 차이**를 고려한, **두 변수 간의 선형 관계** 정도

  1. 타겟변수와 각 설명변수 간에, lag 값마다 상관계수 계산
  2. 가장 높은 상관계수 값을 가지는 lag 선정

[lag = 1]

	Timestamp	LAB_8CHIP_L	8PIC_D4210.PV	8PIC_D4210.PV_lag1
0	2018-11-10 00:00:00	84.599998	111.946831	111.946831
1	2018-11-10 00:01:00	84.599998	112.259820	112.259820
2	2018-11-10 00:02:00	84.599998	109.442167	109.442167
3	2018-11-10 00:03:00	84.599998	106.624023	106.624023
4	2018-11-10 00:04:00	84.599998	103.962240	

상관계수

[lag = 2]

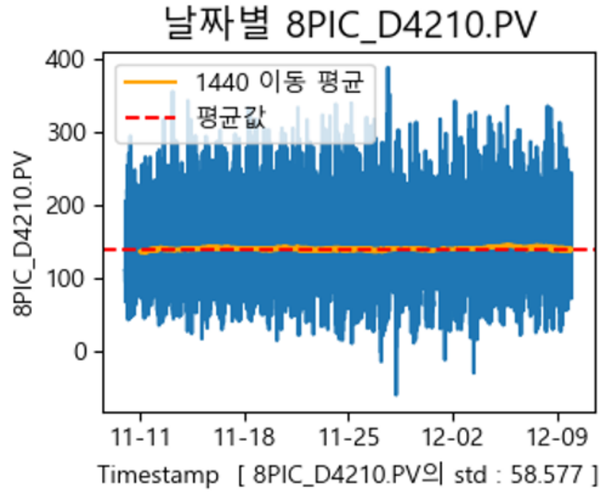
	Timestamp	LAB_8CHIP_L	8PIC_D4210.PV	8PIC_D4210.PV_lag2
0	2018-11-10 00:00:00	84.599998	111.946831	111.946831
1	2018-11-10 00:01:00	84.599998	112.259820	112.259820
2	2018-11-10 00:02:00	84.599998	109.442167	109.442167
3	2018-11-10 00:03:00	84.599998	106.624023	
4	2018-11-10 00:04:00	84.599998	103.962240	

상관계수

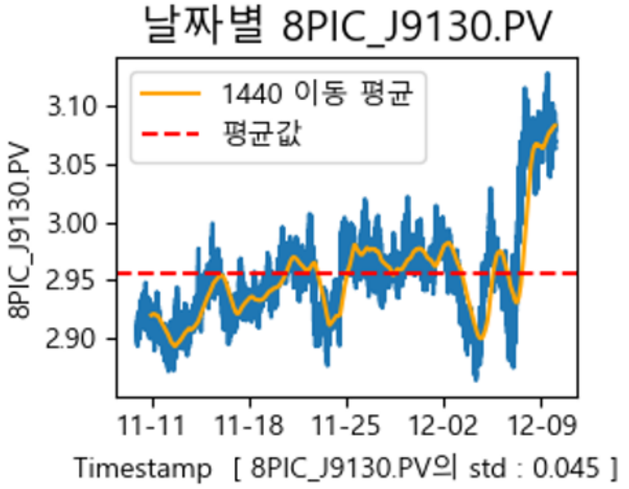
ex) a변수와 타겟변수가 lag 1 에서 가장 높은 상관 계수를 가진다. => 의미) 1 시간 전 a변수 값이 현재 타겟변수 값과 높은 선형 관계를 갖는다.

03 교차 상관 분석 교차 상관 분석 준비

- 교차 상관 분석 가정 : **정상** 데이터



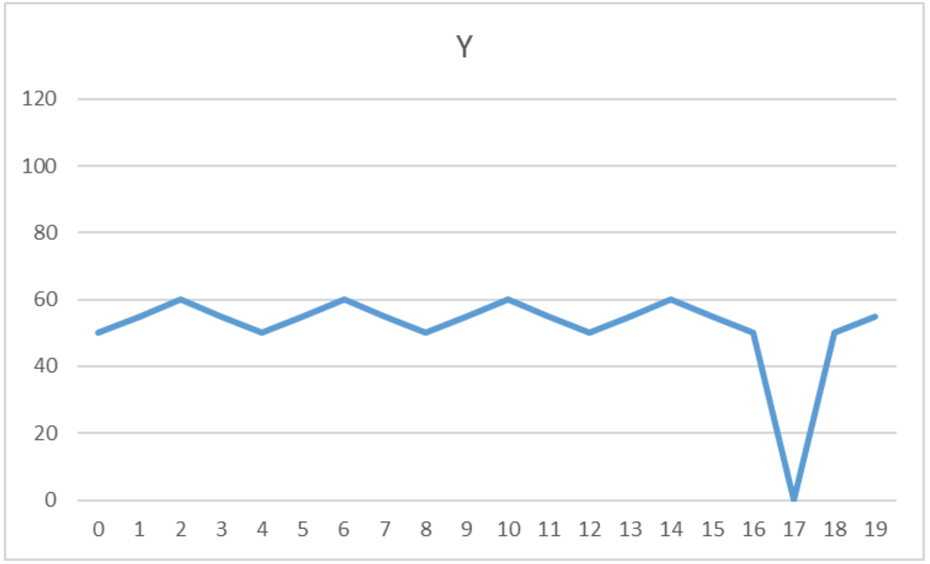
[정상 데이터] 시간 흐름에 따라 평균과 분산이 일정함

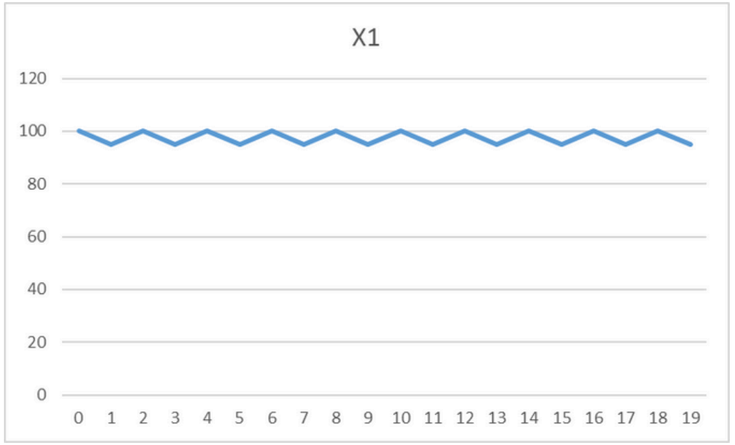


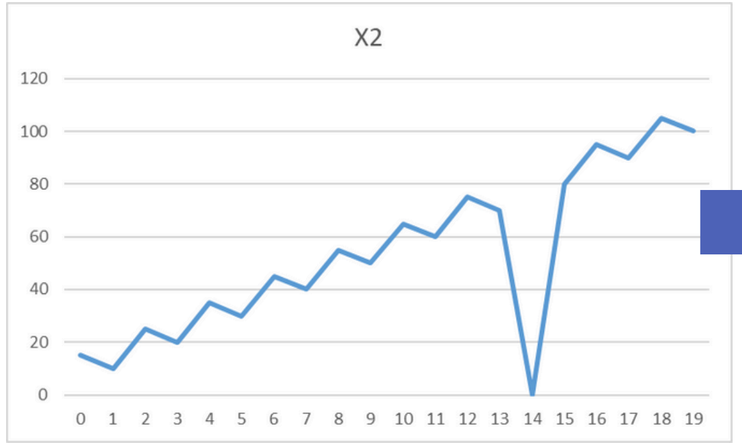
정상화 필요

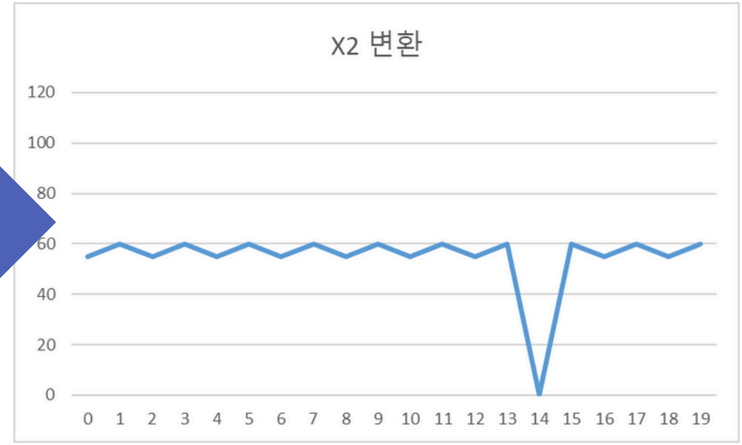
[비정상 데이터] 시간 흐름에 따라 평균과 분산이 달라짐
=> 상관관계가 존재하지 않음에도 불구하고
상관관계가 있는 것처럼 보일 수 있어 정상화 필요

교차 상관 분석 준비


Y
120
100
80
60
40
20
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19


X1
120
100
80
60
40
20
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19


X2
120
100
80
60
40
20
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19


X2 변환
120
100
80
60
40
20
0
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19

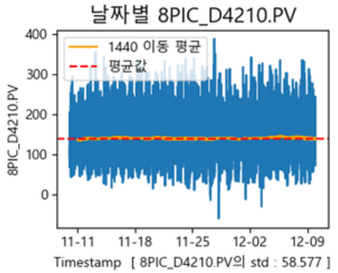
03 교차 상관 분석

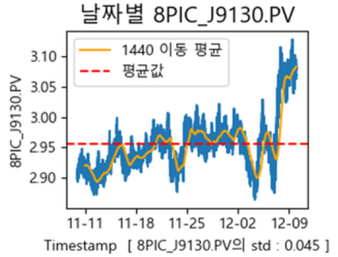
교차 상관 분석 준비

- ADF 검정 : 귀무가설) 해당 변수는 비정상이다
  대립가설) 해당 변수는 정상이다

```
------------------------------
 8PIC_D4210.PV컬럼의 ADF Statistic: -38.948974585078766
8PIC_D4210.PV컬럼의 p-value: 0.0
8PIC_D4210.PV컬럼의 Critical Values:
   1%: -3.4305015746692042
   5%: -2.861606992656292
   10%: -2.5668056580161327
8PIC_D4210.PV컬럼은 정상적이다
------------------------------
 8PIC_J9130.PV컬럼의 ADF Statistic: -2.098215052613026
8PIC_J9130.PV컬럼의 p-value: 0.24523294331753415
8PIC_J9130.PV컬럼의 Critical Values:
   1%: -3.4305015676428203
   5%: -2.861606989550865
   10%: -2.5668056563632007
8PIC_J9130.PV컬럼은 비정상적이다
------------------------------
```

106개 변수가 비정상으로 판단됨





[비정상 데이터를 정상화하는 방법]

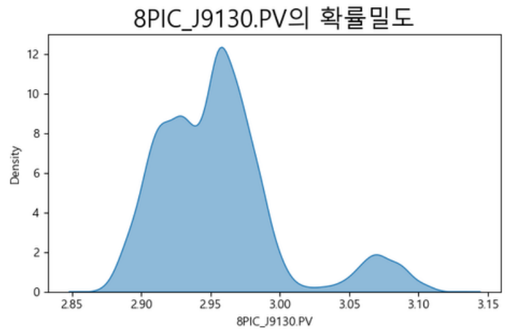
1. 로그 변환
2. 차분

03 교차 상관 분석

교차 상관 분석 준비

1. 로그변환 : 로그 값으로 바꿔줌

확률밀도 확인



오른쪽으로 긴 꼬리를 가짐
-> 로그변환 고려

로그변환 값 계산

	8PIC_J9130.PV	log_8PIC_J9130.PV
0	2.909488	1.363407
1	2.906811	1.362721
2	2.904869	1.362224
3	2.904607	1.362157
4	2.903661	1.361915

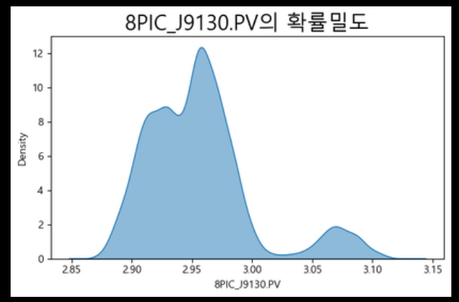
로그

03 교차 상관 분석

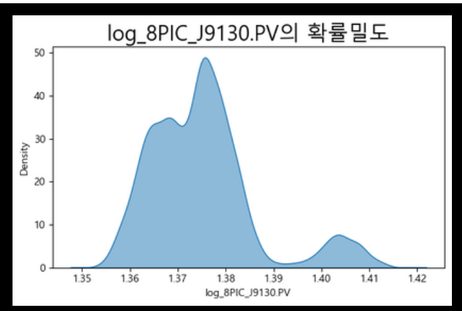
교차 상관 분석 준비

로그 변환후 ADF 검정

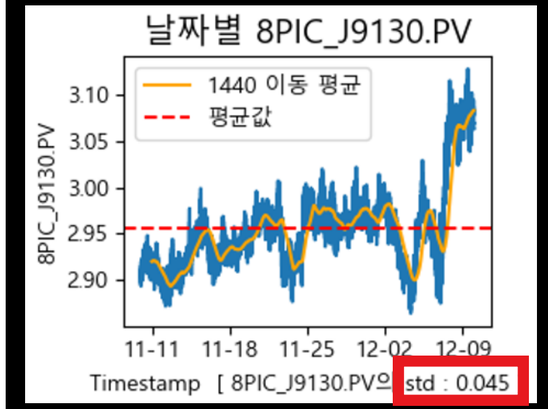
```
------------------------------
 log_8PIC_J9130.PV컬럼의 ADF Statistic: -2.1229063811746807
log_8PIC_J9130.PV컬럼의 p-value: 0.23540413419595851
log_8PIC_J9130.PV컬럼의 Critical Values:
   1%: -3.4305015676428203
   5%: -2.861606989550865
   10%: -2.5668056563632007
log_8PIC_J9130.PV컬럼은 비정상적이다
------------------------------
```



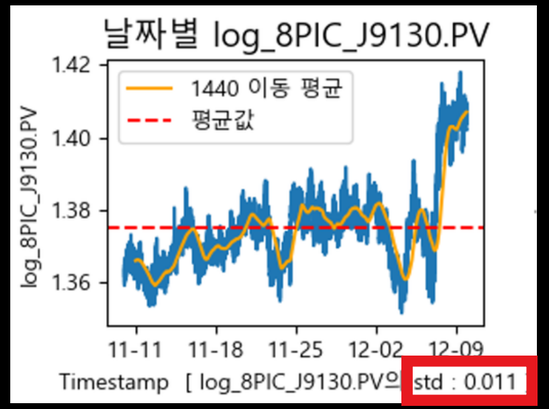
로그



분포가 크게 바뀌지 않음



로그



전체 표준편차가 줄긴 했지만, 큰 변화가 없음

03 교차 상관 분석

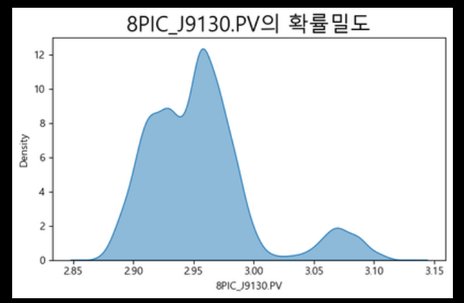
교차 상관 분석 준비

2. 차분 : 이전 값과의 차이

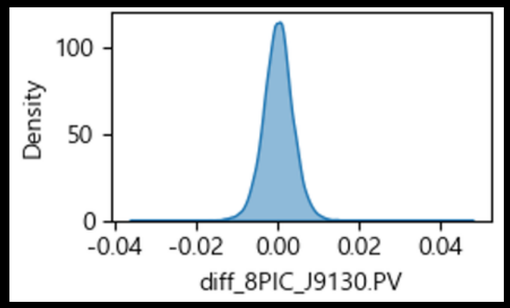
	8PIC_J9130.PV	diff_8PIC_J9130.PV
0	2.909488	NaN
1	2.906811	-0.002677
2	2.904869	-0.001942
3	2.904607	-0.000262
4	2.903661	-0.000945

차분

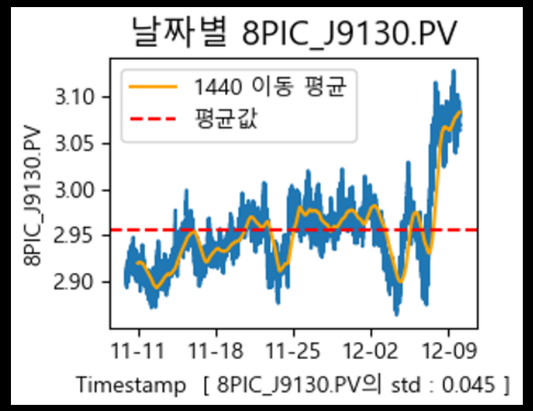
= 2.906811 - 2.909488



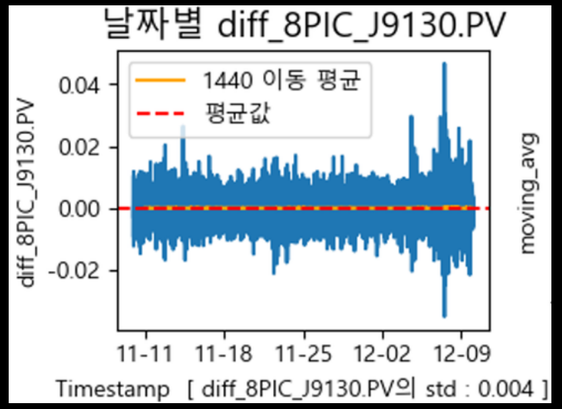
차분



차분 값의 확률 밀도 - 정규분포에 가까워보임



차분



차분후 ADF 검정

```
------------------------------
 diff_8FIC_P4250A.PV컬럼의 ADF Statistic: -46.46947565358915
diff_8FIC_P4250A.PV컬럼의 p-value: 0.0
diff_8FIC_P4250A.PV컬럼의 Critical Values:
   1%: -3.4305015746692042
   5%: -2.861606992656292
   10%: -2.5668056580161327
diff_8FIC_P4250A.PV컬럼은 정상적이다
------------------------------
```

03 교차 상관 분석

연속형 변수

	index	max_abs_corr	original_value	lag
1	8II_E9130B.PV	0.775626	0.775626	1213
2	8TIC_E4312.PV	0.753159	0.753159	22507
3	8PIC_R4330A.PV	0.744877	-0.744877	0
4	8PI_R4330A.PV	0.744763	-0.744763	0
5	8TI_D4250.PV	0.741826	0.741826	0
6	8PI_T4331.PV	0.737694	-0.737694	0
7	8TI_D4240.PV	0.727204	0.727204	0
8	8TI_Z9270H.PV	0.721639	0.721639	5
9	8TI_P4350A.PV	0.713361	0.713361	51
10	8PI_F4272.PV	0.706174	0.706174	0

이산형 변수

	index	max_abs_corr	original_value	lag
1	LAB_8CHIP_B	0.910023	0.910023	0
2	LAB_8CHIP_DEG	0.708623	0.708623	0
3	LAB_8CHIP_COOH	0.491771	-0.491771	0
4	LAB_8CHIP_SIZE	0.486582	-0.486582	2881
5	LAB_8CHIP_TM	0.404032	0.404032	2400
6	LAB_8CHIP_IV	0.382060	-0.382060	480

← 상관관계가 높다고 나오지만, lag 0

← lag 0이 아니지만, 상관관계가 0.48로 낮음

참고) 이산형 변수

```
LAB_8CHIP_IV컬럼의 고유값 개수 : 8
LAB_8CHIP_B컬럼의 고유값 개수 : 11
LAB_8CHIP_DEG컬럼의 고유값 개수 : 5
LAB_8CHIP_COOH컬럼의 고유값 개수 : 3
LAB_8CHIP_TM컬럼의 고유값 개수 : 7
LAB_8CHIP_SIZE컬럼의 고유값 개수 : 2
```

이산형 변수에 대해서는
타겟변수와 상관 관계가
높은 변수가 없다

04 회귀분석

04 회귀분석

상관계수 높은 상위 3개 변수
변수 별 회귀 분석 진행

상위 3개 변수 모두
p_value < 유의수준 0.05 으로
유의미한 변수임을 확인

- lag 1213 고려한 '8II_E9130B.PV' 변수

	coef	std err	t	P>|t|
const	65.4561	0.086	765.269	0.000
8II_E9130B.PV_lag1213	6.1259	0.028	219.899	0.000

R-squared:	0.535
Adj. R-squared:	0.535
F-statistic:	4.836e+04
Prob (F-statistic):	0.00

- lag 22507 고려한 '8TIC_E4312.PV' 변수

	coef	std err	t	P>|t|
const	81.6809	0.010	8011.546	0.000
8TIC_E4312.PV_lag22507	1.3829	0.005	258.352	0.000

R-squared:	0.763
Adj. R-squared:	0.763
F-statistic:	6.675e+04
Prob (F-statistic):	0.00

8TIC_E4312.PV 변수가 다른 변수들 비해 타켓 변수의 변동을 76%로 많이 설명함.

- lag 5 고려한 '8TI_Z9270H.PV' 변수

	coef	std err	t	P>|t|
const	-308.3417	3.934	-78.375	0.000
8TI_Z9270H.PV_lag5	1.4902	0.015	99.794	0.000

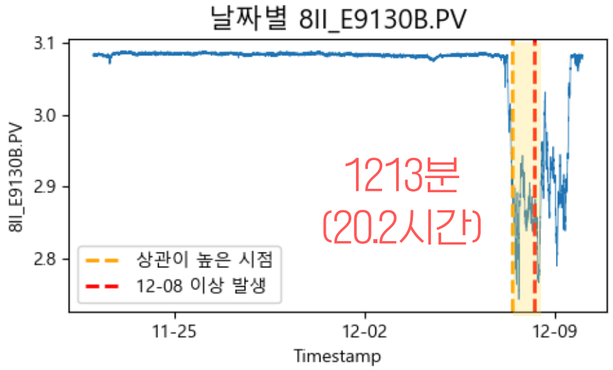
R-squared:	0.187
Adj. R-squared:	0.187
F-statistic:	9959.
Prob (F-statistic):	0.00

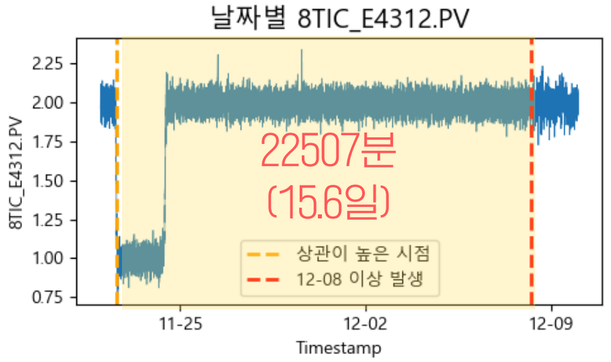
05 결론도출

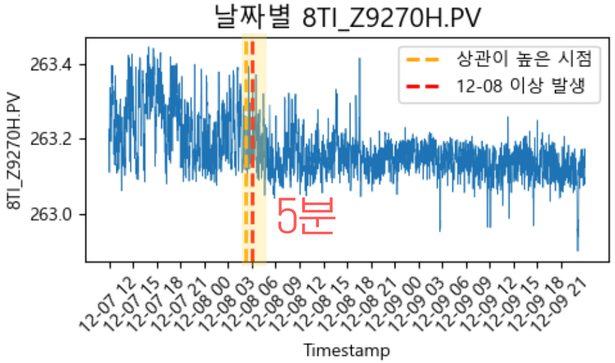
05 결론도출

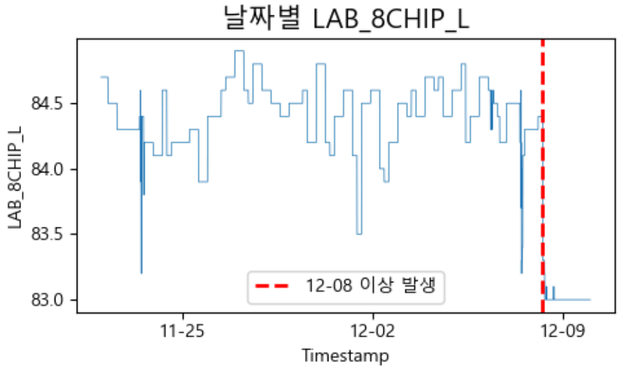
상위 3개 변수

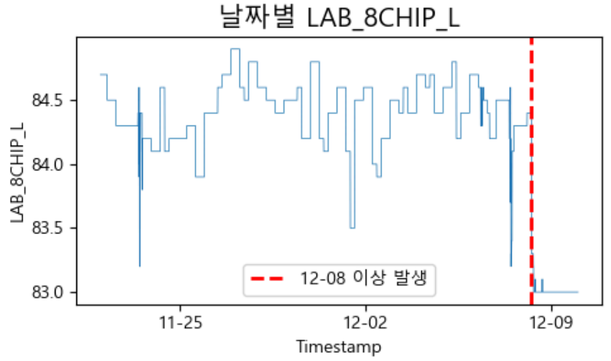
- 노란색 점선 부분에 의해서 이상 현상이 발생한 것이라고 판단
- 해당 lag 마다 모니터링 제안

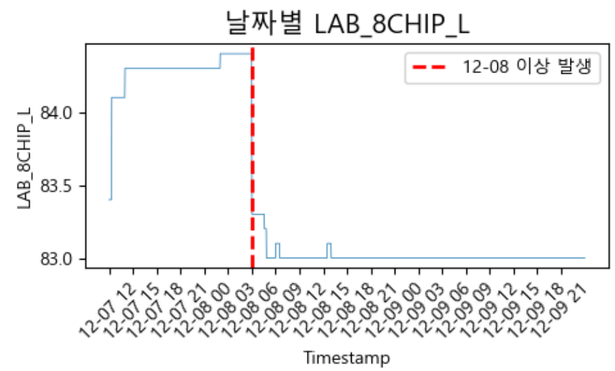












06 Q&A

Q & A

궁금한 점이 있다면 자유롭게 질문해 주세요.

THANK YOU

감사합니다