댐 유입 수량 예측을 통한 최적의 수량 예측 모형 도출

2021 빅콘테스트 퓨처스리그 홍수 ZERO

[빅콘푸로스트]

유현주 (yuhj2049@gmail.com)
서민지 (lyn7666@gmail.com)
정찬웅 (jcw7468@naver.com)

이슈 & 문제점

"기후변화 예측 못한 댐 관리가 작년 홍수피해 키웠다"

[투데이에너지 송명규 기자] 지난 8월 용담댐, 섬진강댐, 합천댐 하류의 홍수피해 원인에 대한 국정감사가 계속되는 가운데 댐을 관리하고 있는 한국수자원공사가 8월 강우량에 대한 부정확한 자체 예측으로 사전 방류 골든타임을 놓쳤다는 지적이 제기됐다.

국회 환경노동위원회 이수진 의원은 한국수자원공사로부터 제출받은 자료를 분석한 결과 이러한 사실이 드러났다고 밝혔다.

'코스핌모형' 예측 엉터리..홍수피해 키워"

올여름 집중호우로 발생한 용담댐 하류지역 홍수 피해의 원인으로 한국수자원공사(이하 수공)가 홍수통제소에 방류 승인 공문을 받을 당시 첨부한 코스핌모형 예측이 엉터리였다는 지적이 나왔다.

"지난해 댐 하류 수해, '기후변화 못 따라간' 댐 운영 등이 원인"

배덕효 한국수자원학회장은 "댐 방류는 댐 초기 수위, 기상예보상 유입량이 얼마나 예측되는지에 따라 종합적으로 결정되는 구조이기 때문에 단순히 방류량만을 갖고 적절성 여부를 답변하기는 곤란하다"면서도 "댐관리규정, 지침·매뉴얼 등에서 '댐 준공 당시' 계획방류량을 현재까지 그대로 유지하는 등 이상기후에 따른 여건 변화를 반영하는 노력이 장기간 부족했다"고 밝혔다.

이슈 & 문제점

"기후변화 예측 못한 댐 관리가 작년 홍수피해 키웠다"

[투데이에너지 송명규 기자] 지난 8월 용담댐, 섬진강댐, 합천댐 하류의 홍수피해 원인에 대한 국정감사가 계속되는 가운데 댐을 관리하고 있는 한국수자원공사가 8일 강우량에 대한 부정확한 자체 예측으로 사전 방류 골든타임을 놓쳤다는 지적이 제기됐다.
국회 환경노동위원회 이수진 의원은 한국수자원공사로부터 제출받은 자료를 분석한 결과 이러한 사실이 드러났다고 밝혔다.

COSFIM 물리적 모형을 사용하여 홍수량 예측에 한계로 인한 홍수 피해 ↑

⇒ AI 기반의 모델이 필요

'코스핌모형' 예측 엉터리 ..홍수피해 키워"

올여름 집중호우로 발생한 용담댐 하류지역 홍수 피해의 원인으로 한국수자원공사(이하 수공)가 홍수통제소에 방류 승인 공문을 받을 당시 첨부한 코스핌모형 예측이 엉터리였다는 지적이 나왔다.

"지난해 댐 하류 수해, '기후변화 못 따라간' 댐 운영 등이 원인"

배덕효 한국수자원학회장은 "댐 방류는 댐 초기 수위, 기상예보상 유입량이 얼마나 예측되는지에 따라 종합적으로 결정되는 구조이기 때문에 단순히 방류량만을 갖고 적절성 여부를 답변하기는 곤란하다"면서도 "댐관리규정, 지침·매뉴얼 등에서 '댐 준공 당시' 계획방류량을 현재까지 그대로 유지하는 등 이상기후에 따른 여건 변화를 반영하는 노력이 장기간 부족했다"고 밝혔다.

CONTENTS

Ⅰ. 데이터 설명

데이터 설명

- **제공데이터**

  : 원본데이터 가공 후

```
data.columns = ['홍수사상번호', '연', '월', '일', '시간', '유입량',
'1_유역평균강수', '1_강우(A지역)', '1_강우(B지역)', '1_강우(C지역)', '1_강우(D지역)', '1_수위(E지역)', '1_수위(D지역)',
'2_유역평균강수', '2_강우(A지역)', '2_강우(B지역)', '2_강우(C지역)', '2_강우(D지역)', '2_수위(E지역)', '2_수위(D지역)',
'3_유역평균강수', '3_강우(A지역)', '3_강우(B지역)', '3_강우(C지역)', '3_강우(D지역)', '3_수위(E지역)', '3_수위(D지역)',
'4_유역평균강수', '4_강우(A지역)', '4_강우(B지역)', '4_강우(C지역)', '4_강우(D지역)', '4_수위(E지역)', '4_수위(D지역)',
'5_유역평균강수', '5_강우(A지역)', '5_강우(B지역)', '5_강우(C지역)', '5_강우(D지역)', '5_수위(E지역)', '5_수위(D지역)',
'6_유역평균강수', '6_강우(A지역)', '6_강우(B지역)', '6_강우(C지역)', '6_강우(D지역)', '6_수위(E지역)', '6_수위(D지역)']
```

	홍수사상번호	연	월	일	시간	유입량	1_유역평균강수	1_강우(A지역)	1_강우(B지역)	1_강우(C지역)	1_강우(D지역)	1_수위(E지역)	1_수위(D지역)	2_유역평균강수	2_강우(A지역)	2_강우(B지역)	2_강우(C지역)	2_강우(D지역)	2_수위(E지역)	2_수위(D지역)
0	1.0	2006.0	7.0	10.0	8.0	189.100000	6.4000	7	7	7	8	2.54	122.56875	6.3000	7	7	7	8	2.54	122.541667
1	1.0	2006.0	7.0	10.0	9.0	216.951962	6.3000	7	8	7	8	2.53	122.56250	6.4000	7	8	7	8	2.53	122.550000
2	1.0	2006.0	7.0	10.0	10.0	251.424419	6.4000	7	9	7	8	2.53	122.55625	7.3000	7	9	7	8	2.53	122.558333
3	1.0	2006.0	7.0	10.0	11.0	302.812199	7.3000	7	10	7	8	2.53	122.55625	8.2000	7	10	8	8	2.53	122.566667
4	1.0	2006.0	7.0	10.0	12.0	384.783406	8.2000	7	12	8	10	2.53	122.55625	11.3000	9	12	10	10	2.53	122.575000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3046	26.0	2018.0	7.0	7.0	17.0	NaN	2.3689	1	0	0	0	3.16	129.99375	2.3689	1	0	0	0	3.16	130.016667
3047	26.0	2018.0	7.0	7.0	18.0	NaN	2.3689	1	0	0	0	3.15	130.00625	2.3689	1	0	0	0	3.15	130.025000
3048	26.0	2018.0	7.0	7.0	19.0	NaN	2.3689	1	0	0	0	3.13	130.01250	2.3689	1	0	0	0	3.13	130.025000
3049	26.0	2018.0	7.0	7.0	20.0	NaN	2.3689	1	0	0	0	3.11	130.01875	2.3689	1	0	0	0	3.11	130.025000
3050	26.0	2018.0	7.0	7.0	21.0	NaN	2.3689	1	0	0	0	3.10	130.01875	2.3689	1	0	0	0	3.10	130.025000

3051 rows × 48 columns

데이터 설명

- **외부 데이터**
  : 기상청 데이터

연	월	1_강우(A지역)	1_강우(B지역)	1_강우(C지역)	1_강우(D지역)
2006	7	251	367	253	262
2007	8	81	89	107	235
	9	127	116	98	116
2008	7	275	220	159	177
2009	7	221	126	117	125
	8	156	98	53	85
2010	9	165	146	105	179
2011	6	142	76	102	91
	7	171	148	121	136
	8	79	112	159	222
2012	7	164	191	163	184
	8	74	90	106	89
	9	55	47	106	81
2013	7	329	209	58	140
2017	7	293	202	166	119
2018	7	109	124	146	156

- 주어진 데이터의 2013년 7월 15일의 강우량과 춘천의 강우량이 유사함

	홍수사상번호	연	월	일	1_강우(A지역)	1_강우(B지역)	1_강우(C지역)	1_강우(D지역)
2479	22	2013	7	15	329	208	50	139

7.11~7.15,7.18 호우	2013.7.11~7.15,7.18	경기(가평), 강원(춘천, 홍천, 평창, 인제)	459mm(춘천)

데이터 설명

- **외부 데이터**
  : 기상청 데이터

	일시	기온	풍속	습도
0	2006-07-10 08:00	24.3	2.0	85.0
1	2006-07-10 09:00	24.6	1.5	83.0
2	2006-07-10 10:00	25.4	1.6	79.0
3	2006-07-10 11:00	25.4	0.9	82.0
4	2006-07-10 12:00	25.3	0.8	81.0
...	...	...	...	...
3046	2018-07-07 17:00	23.5	3.6	56.0
3047	2018-07-07 18:00	23.2	3.4	56.0
3048	2018-07-07 19:00	22.0	3.5	56.0
3049	2018-07-07 20:00	20.0	2.3	61.0
3050	2018-07-07 21:00	18.9	3.0	64.0

3051 rows × 4 columns

춘천 지역 근방의 댐이라고 가정
⇨ 2003 ~ 2018년도의 **기상청 데이터**를 사용

Ⅱ. 탐색적 데이터 설명(EDA)

탐색적 데이터 분석 (EDA)

- 데이터 형태

데이터 칼럼

	홍수사상번호	연	월	일	시간	유입량	1_유역평균강수	1_강우(A지역)	1_강우(B지역)	1_강우(C지역)	...	6_강우(A지역)	6_강우(B지역)	6_강우(C지역)	6_강우(D지역)	6_수위(E지역)	6_수위(D지역)	일시	기온	풍속	습도
0	1.0	2006.0	7.0	10.0	8.0	189.100000	6.4000	7	7	7	...	7	7	8	8	2.54	122.610	2006-07-10 08:00	24.3	2.0	85.0
1	1.0	2006.0	7.0	10.0	9.0	216.951962	6.3000	7	8	7	...	7	8	10	10	2.53	122.600	2006-07-10 09:00	24.6	1.5	83.0
2	1.0	2006.0	7.0	10.0	10.0	251.424419	6.4000	7	9	7	...	7	9	10	11	2.53	122.590	2006-07-10 10:00	25.4	1.6	79.0
3	1.0	2006.0	7.0	10.0	11.0	302.812199	7.3000	7	10	7	...	9	10	15	14	2.53	122.585	2006-07-10 11:00	25.4	0.9	82.0
4	1.0	2006.0	7.0	10.0	12.0	384.783406	8.2000	7	12	8	...	12	12	18	16	2.53	122.575	2006-07-10 12:00	25.3	0.8	81.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3046	26.0	2018.0	7.0	7.0	17.0	NaN	2.3689	1	0	0	...	1	0	0	0	3.16	129.950	2018-07-07 17:00	23.5	3.6	56.0
3047	26.0	2018.0	7.0	7.0	18.0	NaN	2.3689	1	0	0	...	1	0	0	0	3.15	129.970	2018-07-07 18:00	23.2	3.4	56.0
3048	26.0	2018.0	7.0	7.0	19.0	NaN	2.3689	1	0	0	...	1	0	0	0	3.13	129.980	2018-07-07 19:00	22.0	3.5	56.0
3049	26.0	2018.0	7.0	7.0	20.0	NaN	2.3689	1	0	0	...	1	0	0	0	3.11	129.990	2018-07-07 20:00	20.0	2.3	61.0
3050	26.0	2018.0	7.0	7.0	21.0	NaN	2.3689	1	0	0	...	1	0	0	0	3.10	130.000	2018-07-07 21:00	18.9	3.0	64.0

3051 rows × 52 columns

데이터 형태

집단(1~6) 별 데이터

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3051 entries, 0 to 3050
Data columns (total 52 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   홍수사상번호      3051 non-null   float64
 1   연            3051 non-null   float64
 2   월            3051 non-null   float64
 3   일            3051 non-null   float64
 4   시간           3051 non-null   float64
 5   유입량          2891 non-null   float64
 6   1_유역평균강수     3051 non-null   float64
 7   1_강우(A지역)    3051 non-null   int64
 8   1_강우(B지역)    3051 non-null   int64
 9   1_강우(C지역)    3051 non-null   int64
 10  1_강우(D지역)    3051 non-null   int64
 11  1_수위(E지역)    3051 non-null   float64
 12  1_수위(D지역)    3051 non-null   float64
 13  2_유역평균강수     3051 non-null   float64
 14  2_강우(A지역)    3051 non-null   int64
 15  2_강우(B지역)    3051 non-null   int64
 16  2_강우(C지역)    3051 non-null   int64
 17  2_강우(D지역)    3051 non-null   int64
 18  2_수위(E지역)    3051 non-null   float64
 19  2_수위(D지역)    3051 non-null   float64
 20  3_유역평균강수     3051 non-null   float64
 21  3_강우(A지역)    3051 non-null   int64
 22  3_강우(B지역)    3051 non-null   int64
 23  3_강우(C지역)    3051 non-null   int64
 24  3_강우(D지역)    3051 non-null   int64
 25  3_수위(E지역)    3051 non-null   float64
 26  3_수위(D지역)    3051 non-null   float64
 27  4_유역평균강수     3051 non-null   float64
 28  4_강우(A지역)    3051 non-null   int64
 29  4_강우(B지역)    3051 non-null   int64
 30  4_강우(C지역)    3051 non-null   int64
 31  4_강우(D지역)    3051 non-null   int64
 32  4_수위(E지역)    3051 non-null   float64
 33  4_수위(D지역)    3051 non-null   float64
```

데이터 타입

탐색적 데이터 분석 (EDA)

▪ 기본 데이터 기초 통계량

- 유입량 : 75%에 해당하는 값과 MAX 값의 차이가 큼

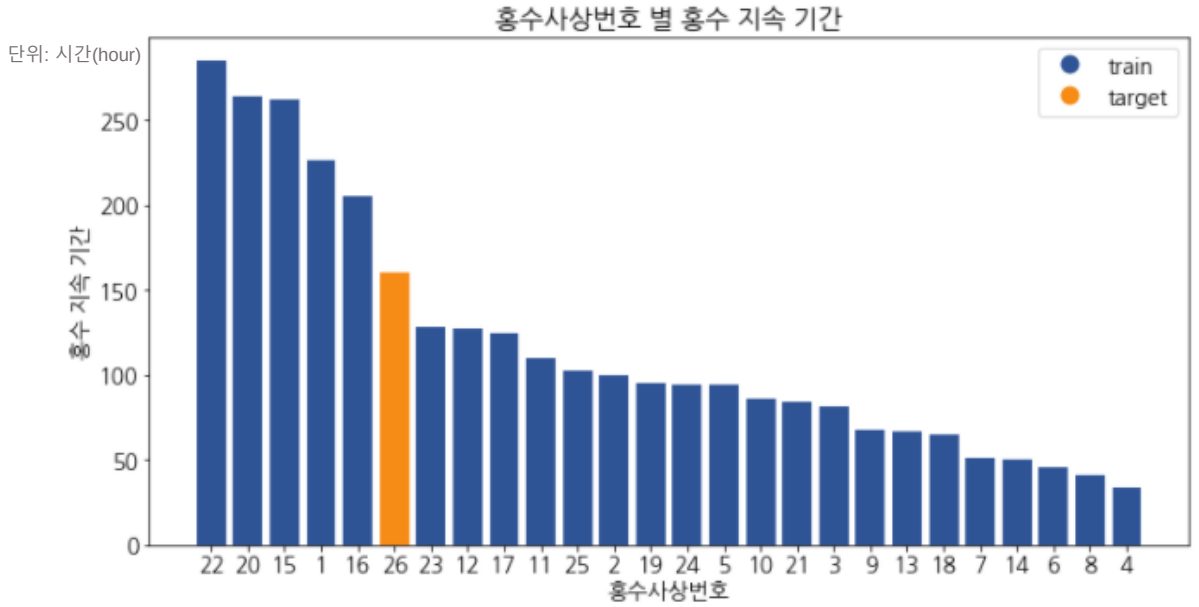
	count	mean	std	min	25%	50%	75%	max
유입량	2891.0	1746.902717	2181.790290	3.541153	384.762514	1065.549253	2213.014288	21504.402260
1_유역평균강수	3051.0	67.938486	59.359959	0.000000	20.500000	57.198300	103.226400	328.400000
1_강우(A지역)	3051.0	78.058669	71.874728	0.000000	22.000000	55.000000	121.000000	329.000000
1_강우(B지역)	3051.0	64.361849	63.369565	0.000000	13.000000	52.000000	91.000000	367.000000
1_강우(C지역)	3051.0	39.822353	46.097700	0.000000	3.000000	22.000000	65.000000	253.000000
1_강우(D지역)	3051.0	53.558178	56.327987	0.000000	7.000000	35.000000	85.000000	262.000000
1_수위(E지역)	3051.0	4.564936	2.375226	1.070000	2.810000	4.080000	5.580000	16.720000
1_수위(D지역)	3051.0	131.539823	6.061602	118.700000	127.331250	133.012500	135.175000	143.893750
2_유역평균강수	3051.0	68.420307	59.315391	0.000000	20.917250	58.500000	103.413650	328.400000
2_강우(A지역)	3051.0	81.216978	72.927209	0.000000	23.000000	57.000000	123.000000	337.000000
2_강우(B지역)	3051.0	63.484104	62.906248	0.000000	13.000000	52.000000	90.000000	367.000000
2_강우(C지역)	3051.0	39.237955	45.493394	0.000000	3.000000	21.000000	64.000000	251.000000
2_강우(D지역)	3051.0	53.558178	56.327987	0.000000	7.000000	35.000000	85.000000	262.000000
2_수위(E지역)	3051.0	4.564936	2.375226	1.070000	2.810000	4.080000	5.580000	16.720000
2_수위(D지역)	3051.0	131.611259	6.060190	118.700000	127.333333	133.050000	135.200000	143.966667

탐색적 데이터 분석 (EDA)

▪ 홍수사상번호

- 홍수를 구분하는 고유번호로 홍수사상별로 지속기간(단위:시간)이 다름
- 홍수사상번호 22번의 기간이 가장 길고, 4번이 가장 짧음

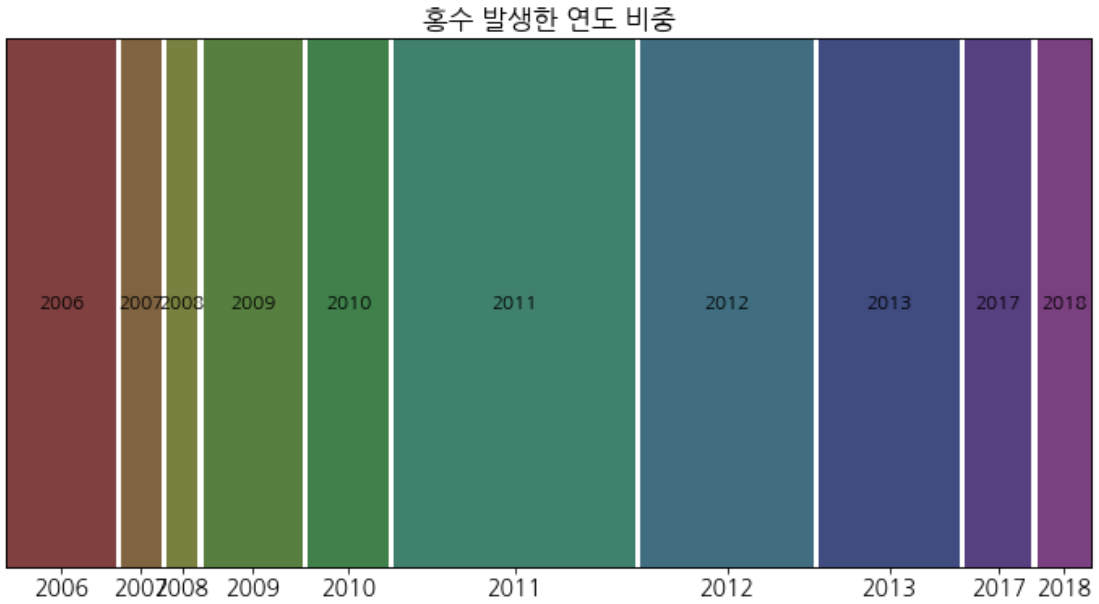
	홍수사상번호	연	월	일	시간	유입량	1_유역평균강수	1_강우(A지역)	1_강우(B지역)	1_강우(C지역)	...
224	1.0	2006.0	7.0	19.0	16.0	3373.123471	105.3	125	88	72	...
225	1.0	2006.0	7.0	19.0	17.0	3285.961383	103.2	121	86	71	...
226	2.0	2006.0	7.0	25.0	24.0	323.993267	0.0	0	1	0	...
227	2.0	2006.0	7.0	26.0	1.0	323.154138	0.0	0	2	0	...

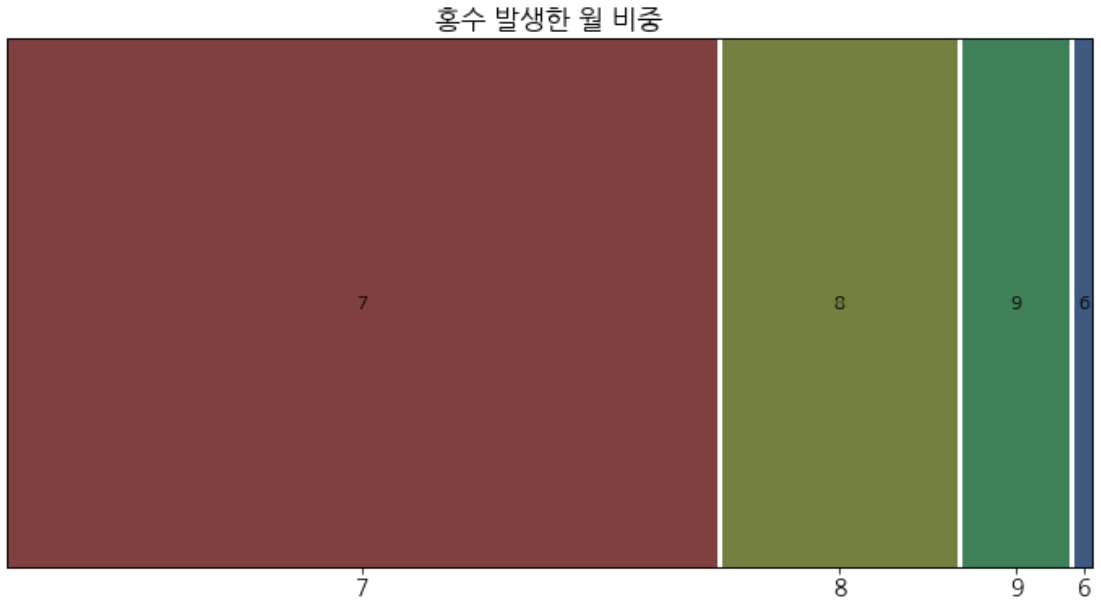


탐색적 데이터 분석 (EDA)

▪ 연, 월, 일, 시간

- 홍수가 발생한 날짜 및 시간 데이터
- 홍수사상별로 홍수가 발생한 시기는 모두 다름
- 홍수는 2011년(연도 비중), 7월(월 비중)에 홍수의 발생빈도가 많음

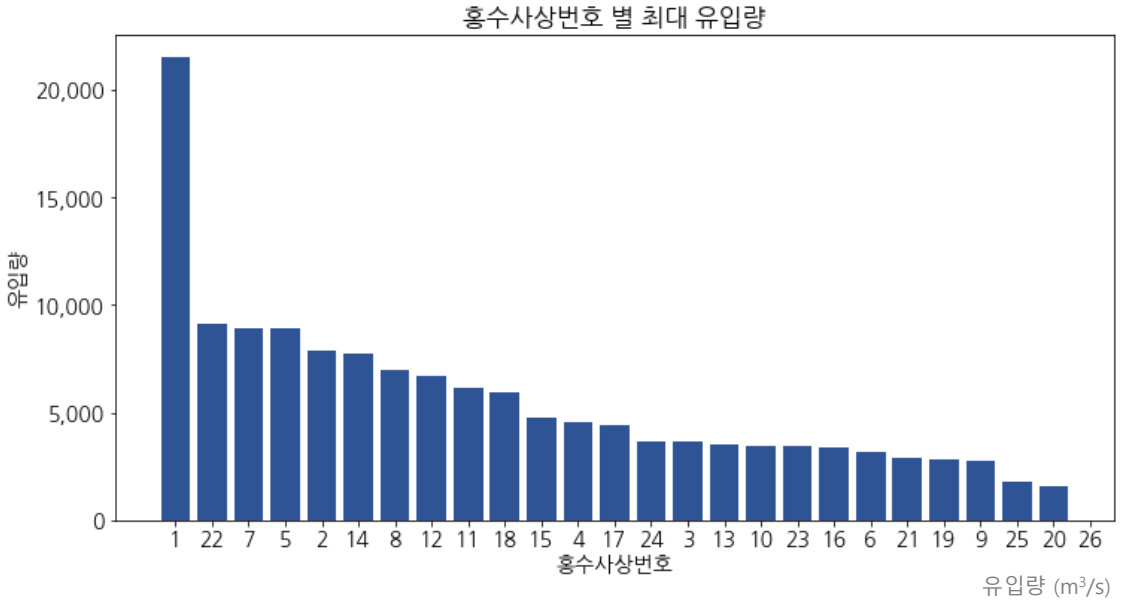


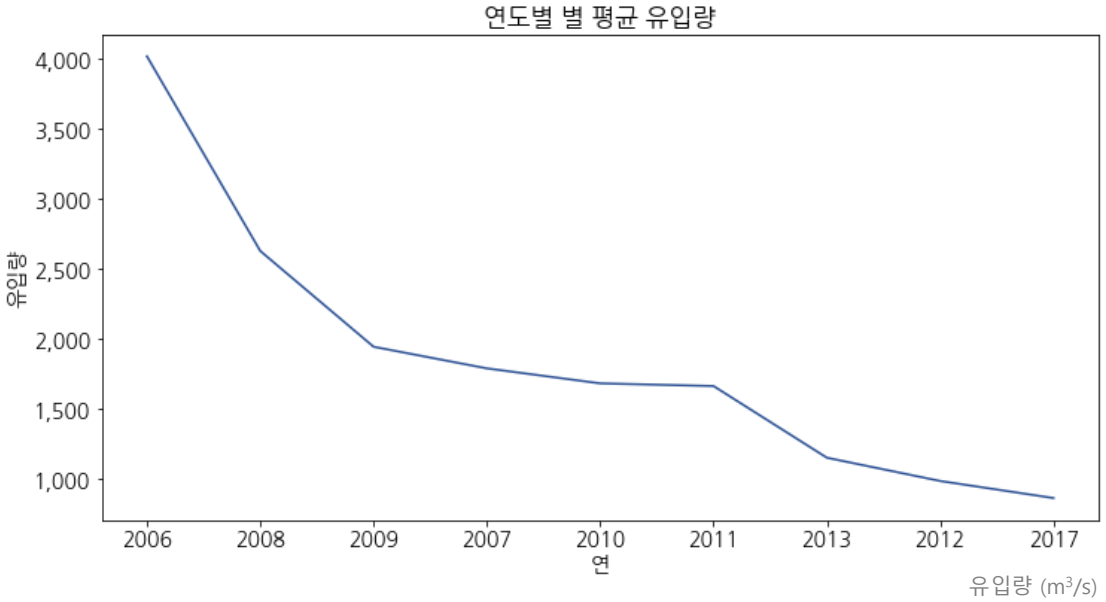


탐색적 데이터 분석 (EDA)

▪ 유입량

- K-댐에 흘러 들어오는 유량으로 예측하고자 하는 값
- 홍수사상 1번은 20000 m^3/s 이상의 최대값
- 1번을 제외한 홍수사상들의 경우 최댓값이 10000 m^3/s 미만
- 연도별 평균 유입량은 줄어드는 추세를 보임

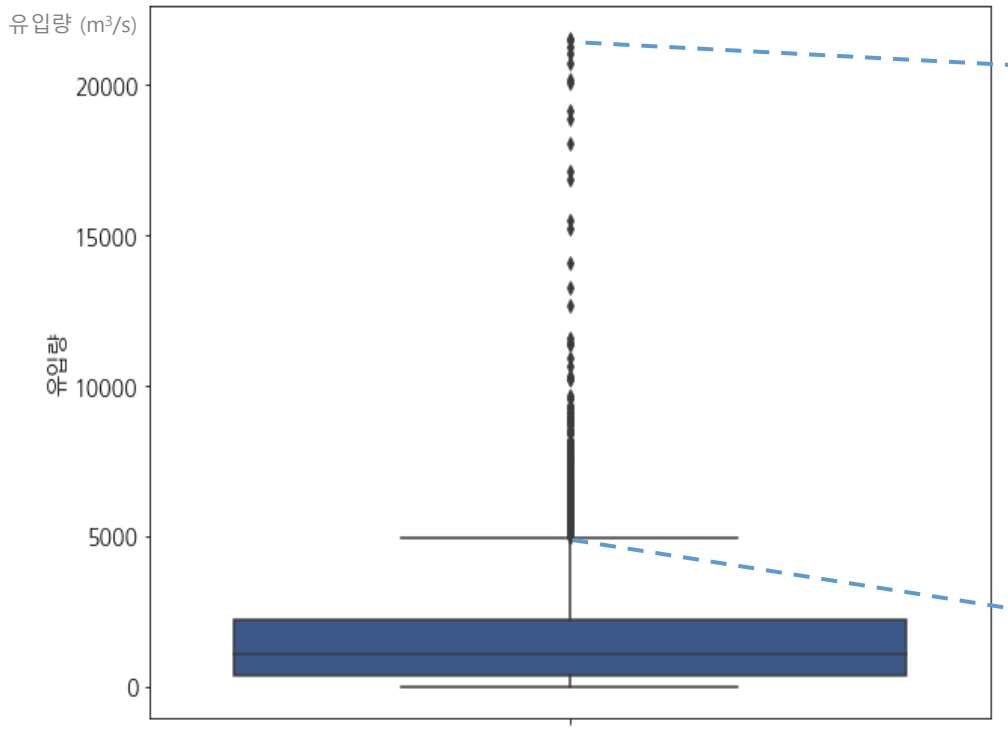




탐색적 데이터 분석 (EDA)

▪ 유입량

- 유입량의 boxplot을 통해, outlier에 해당하는 대부분의 값이 홍수사상번호 1번임을 확인

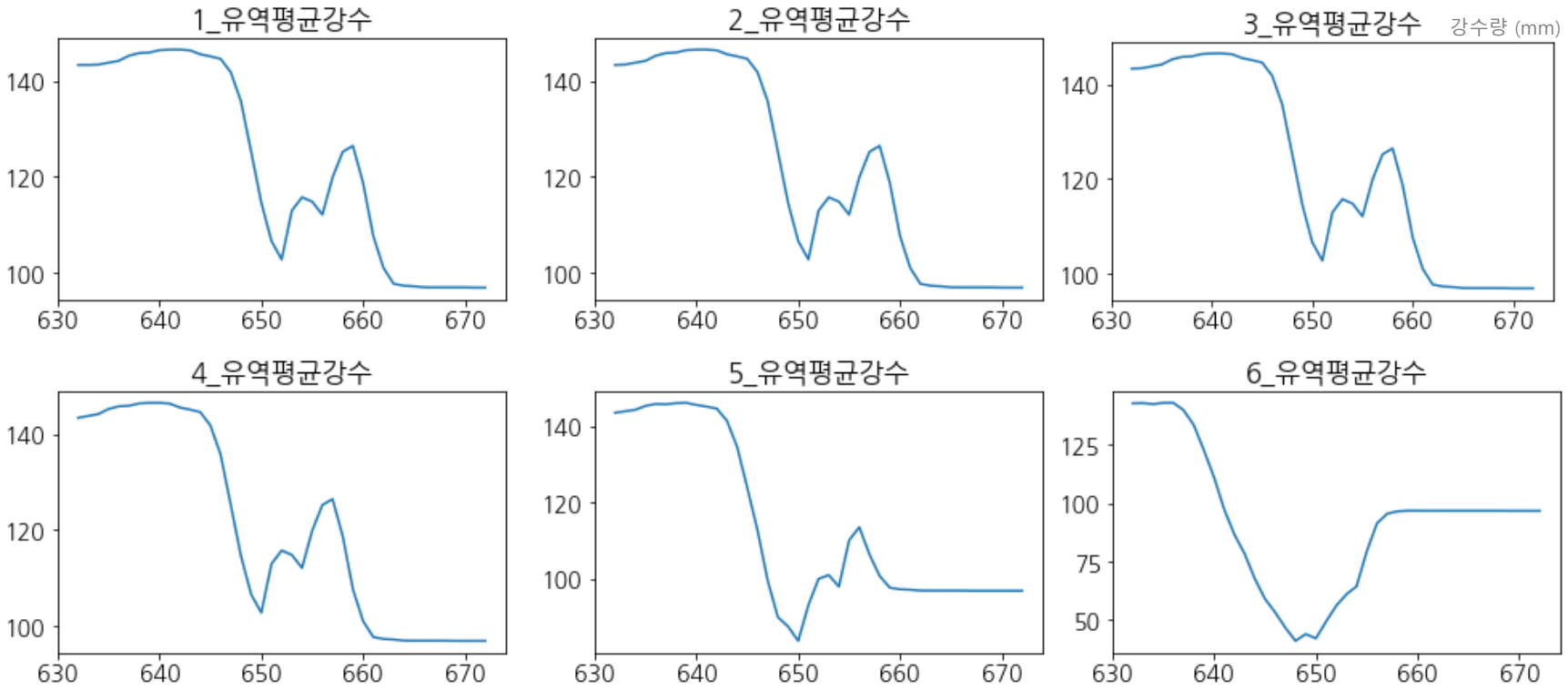


홍수사상번호	유입량
1.0	67
2.0	31
5.0	21
7.0	15
8.0	12
11.0	9
12.0	9
14.0	15
18.0	8
22.0	14

탐색적 데이터 분석 (EDA)

▪ 유역평균강수

- 전체 유역의 평균 누적 강수량 데이터
- 홍수사상별로 6개 데이터 집단의 유역평균강수의 분포가 유사하나 약간의 차이를 보임

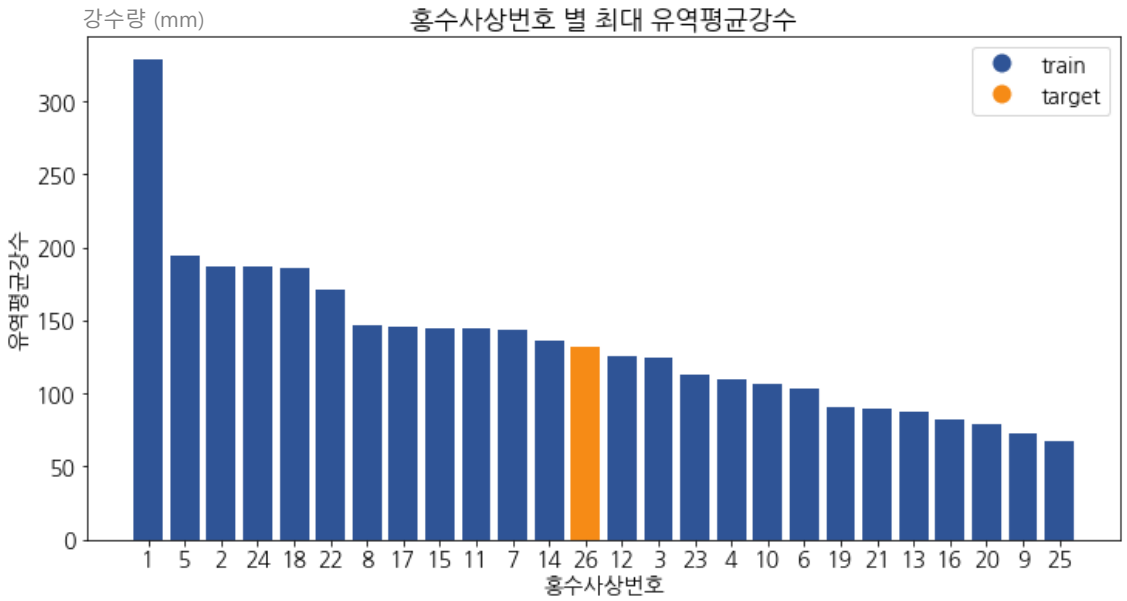


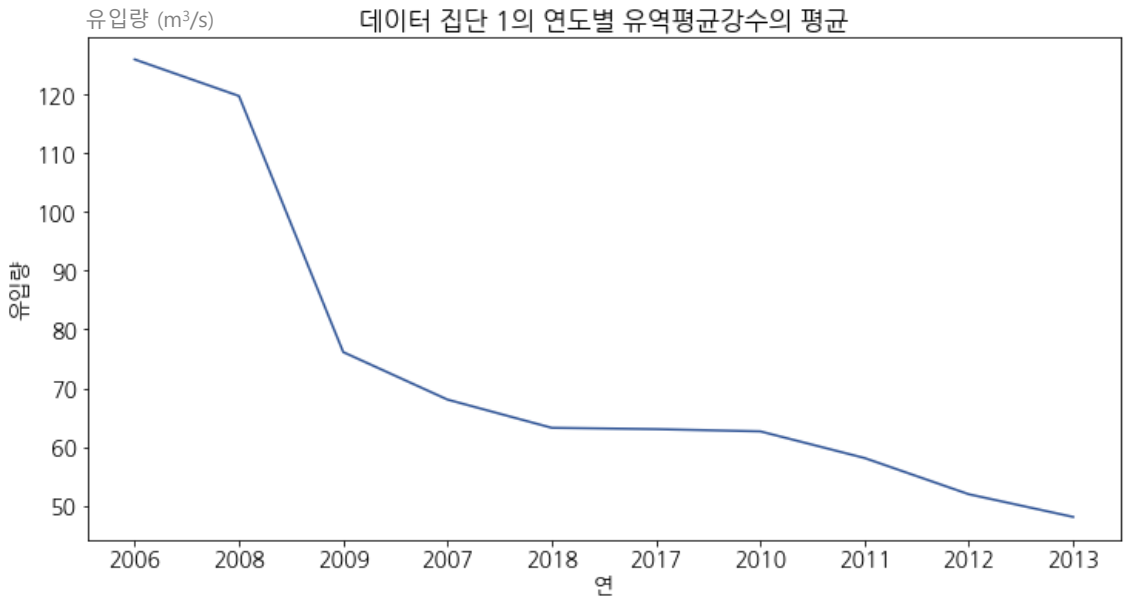
〈홍수사상번호 8번 기준〉

탐색적 데이터 분석 (EDA)

▪ 유역평균강수

- 홍수사상별로 최대 유역평균강수를 확인하였을 때 홍수사상 1번이 최대임을 확인
- 연도별 유역평균강수의 분포는 점점 줄어드는 추세

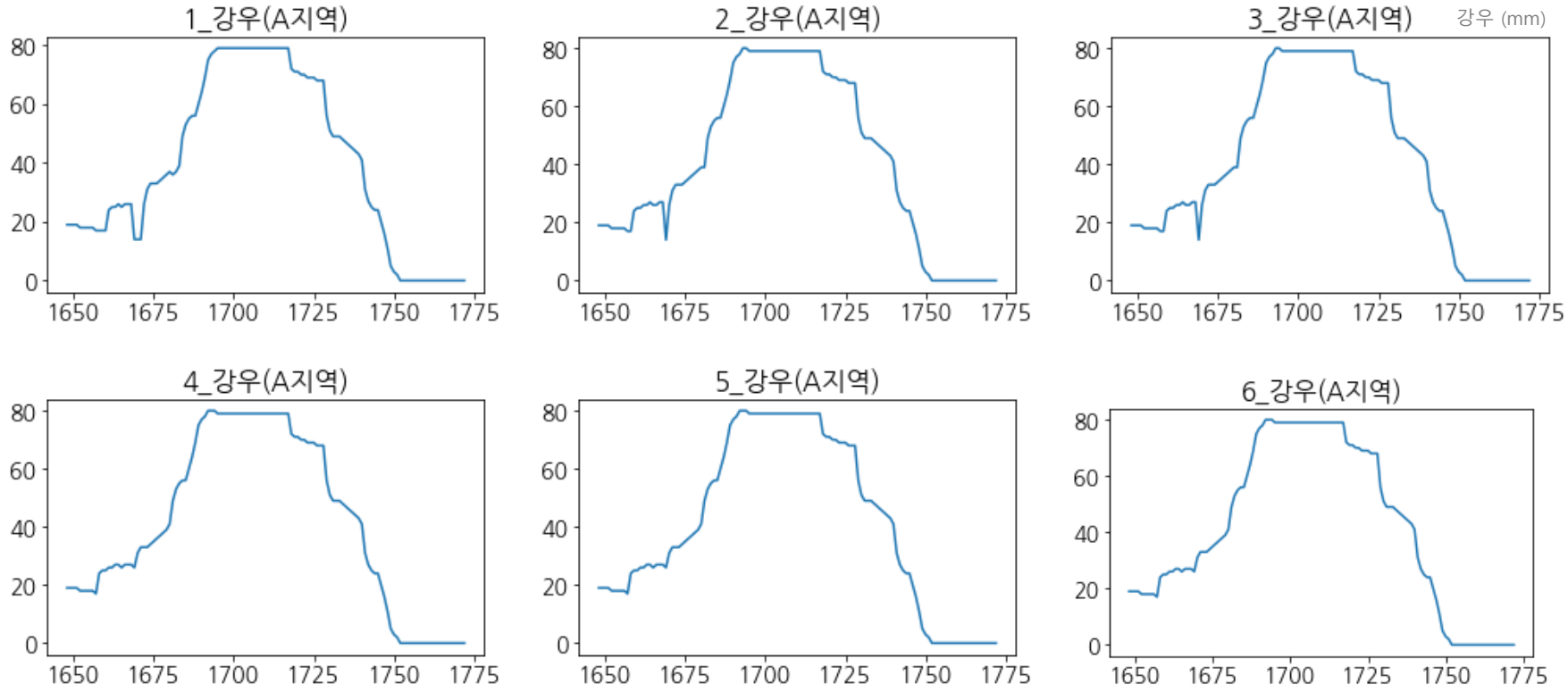




탐색적 데이터 분석 (EDA)

▪ 강우(A지역)

- A 관측소의 누적강수량 데이터
- 홍수사상별로 6개 데이터 집단의 강우(A지역)의 분포가 유사하나 약간의 차이를 보임

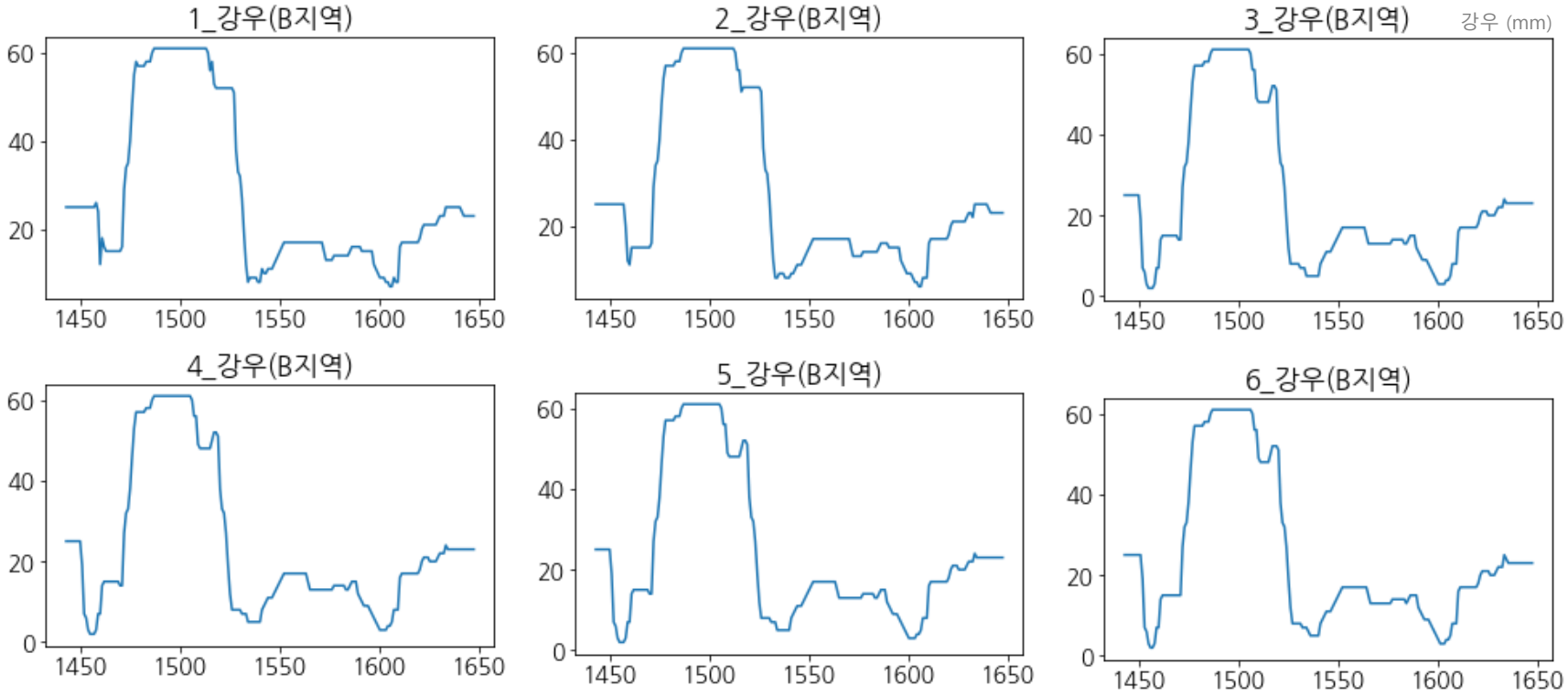


〈홍수사상번호 17번 기준〉

탐색적 데이터 분석 (EDA)

▪ 강우(B지역)

- B 관측소의 누적강수량 데이터
- 홍수사상별로 6개 데이터 집단의 강우(B지역)의 분포가 유사하나 약간의 차이를 보임

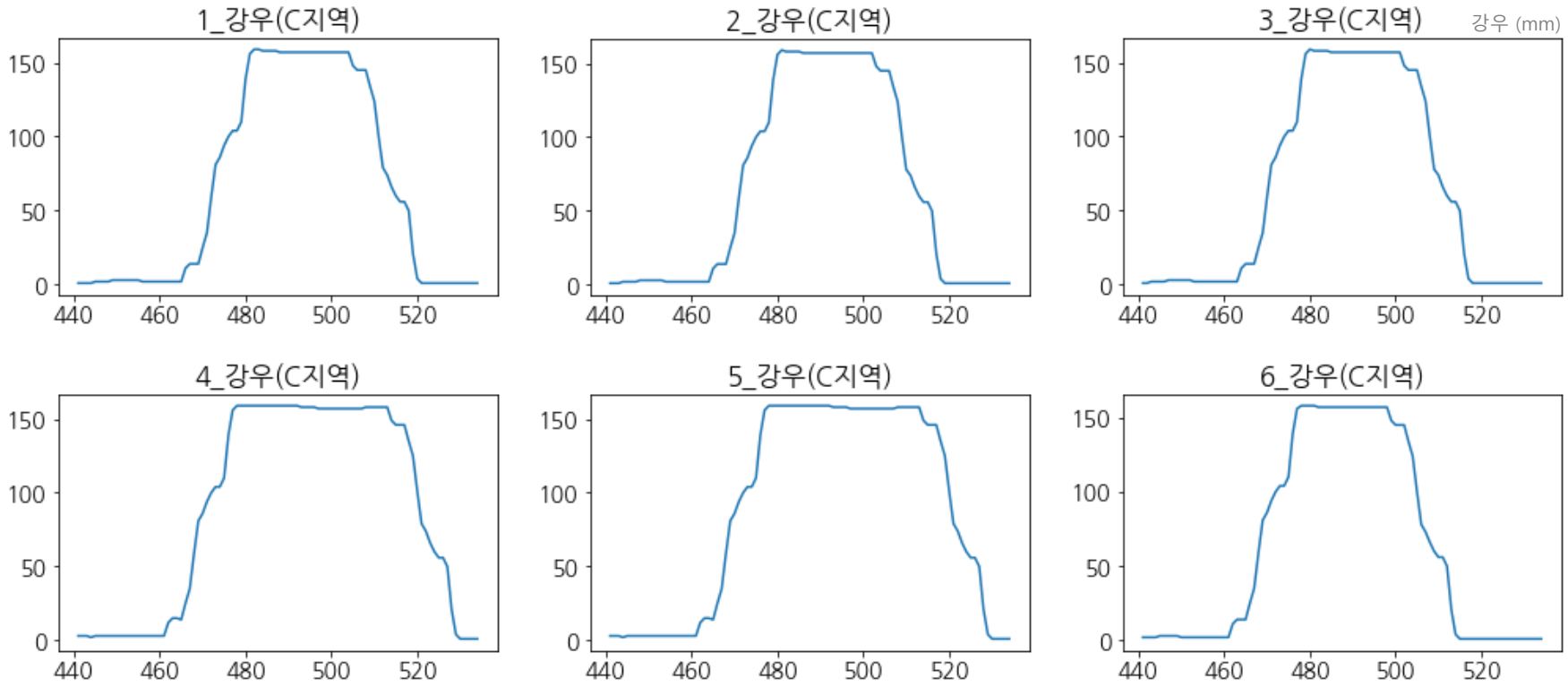


〈홍수사상번호 16번 기준〉

탐색적 데이터 분석 (EDA)

▪ 강우(C지역)

- C 관측소의 누적강수량 데이터
- 홍수사상별로 6개 데이터 집단의 강우(C지역)의 분포가 유사하나 약간의 차이를 보임

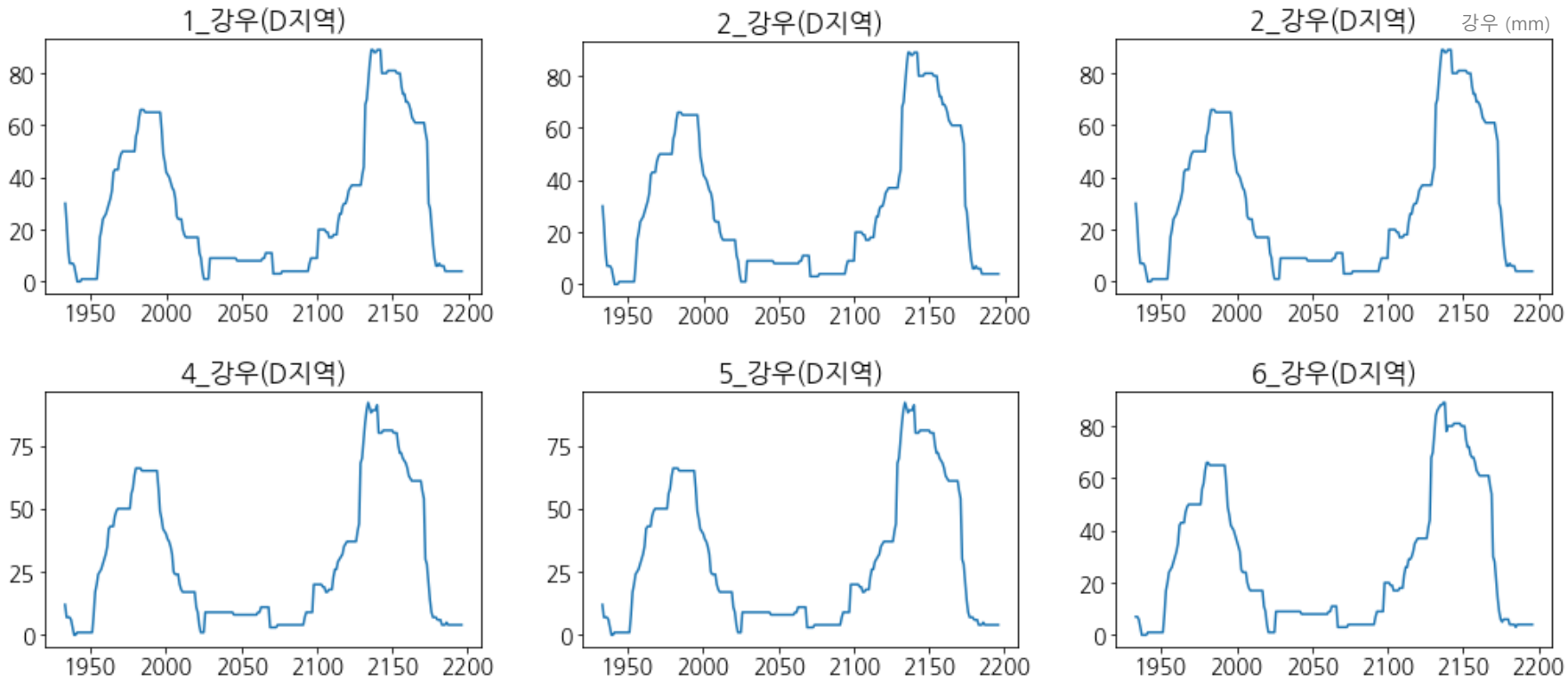


〈홍수사상번호 5번 기준〉

탐색적 데이터 분석 (EDA)

▪ 강우(D지역)

- D 관측소의 누적강수량 데이터
- 홍수사상별로 6개 데이터 집단의 강우(D지역)의 분포가 유사하나 약간의 차이를 보임



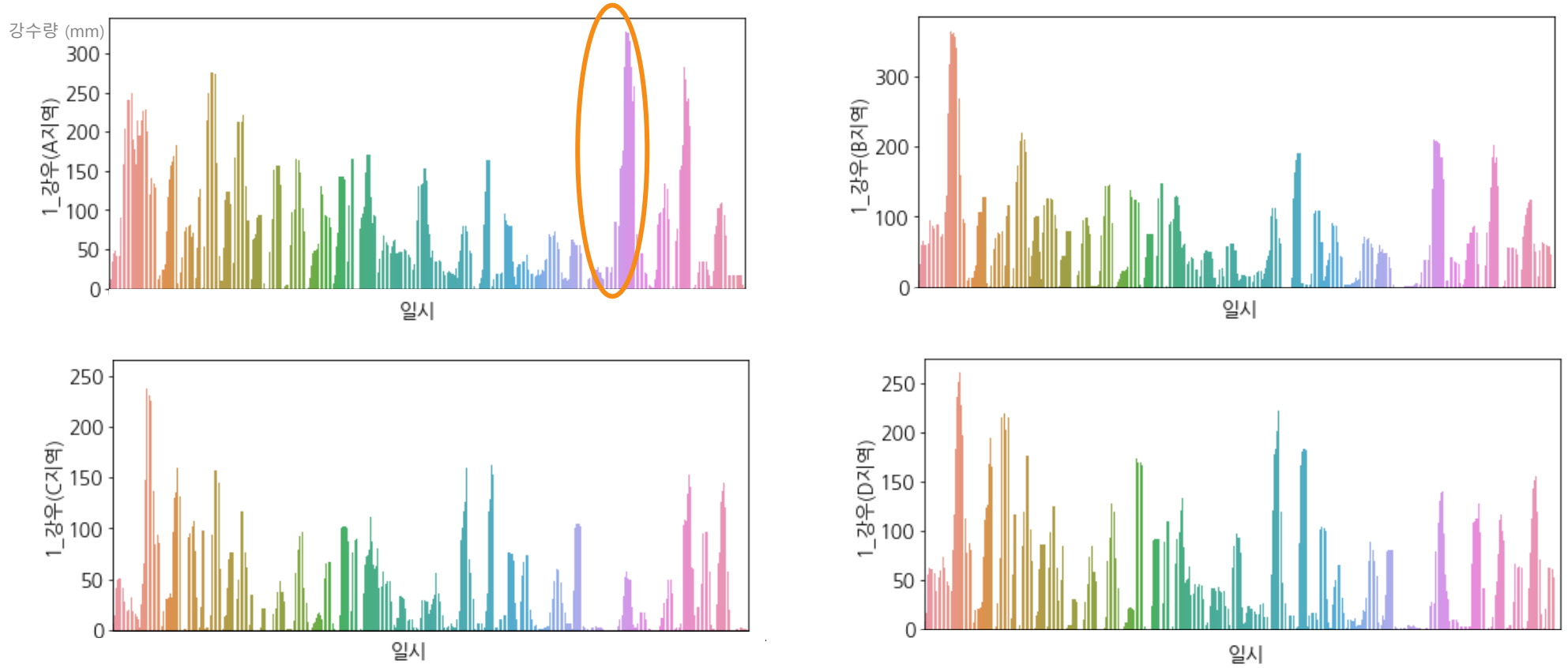
〈홍수사상번호 20번 기준〉

탐색적 데이터 분석 (EDA)

▪ 강우(A, B, C, D지역)

- 데이터 집단 1에 대하여 강우 A, B, C, D 지역의 분포를 비교하면 다음과 같음
- 4개의 분포 중 강우 (A지역)의 분포는 다른 지역에 비해 2013년도에 높은 강우량이 나타남

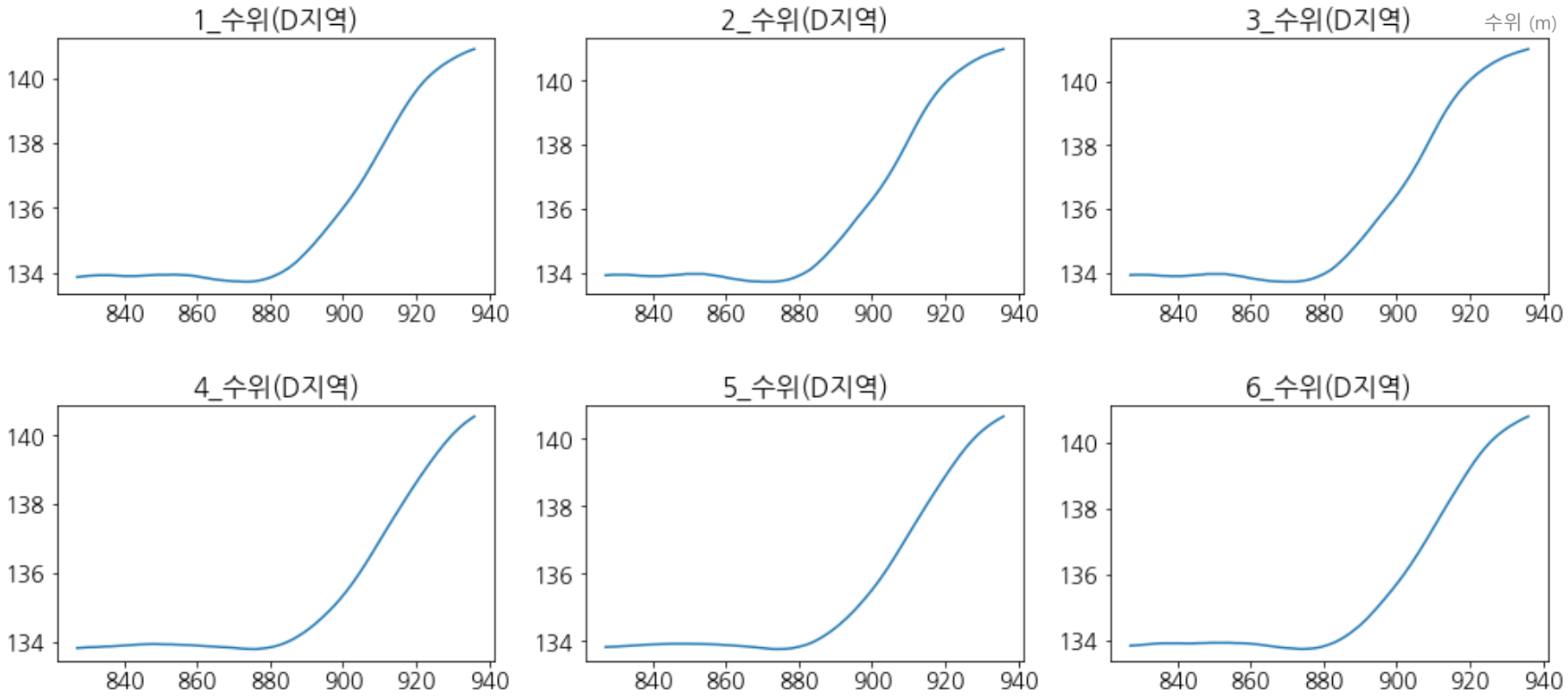
  ⇒ 관측소 A는 2013년에 홍수 피해를 겪은 곳이라고 추측



탐색적 데이터 분석 (EDA)

▪ 수위(D지역)

- D 관측소의 수위 데이터
- 홍수사상별로 6개 데이터 집단의 강우(D지역)의 분포가 유사하나 약간의 차이를 보임

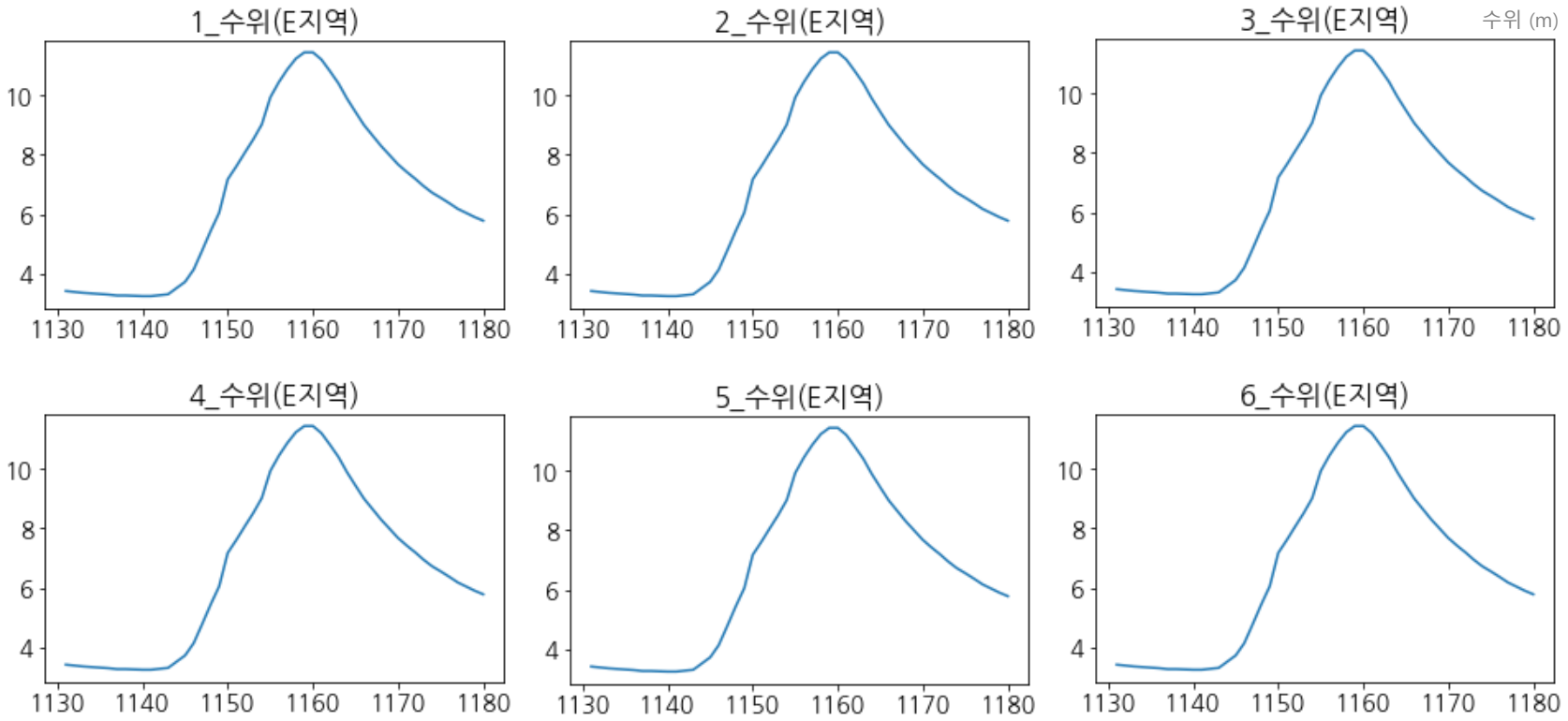


〈홍수사상번호 11번 기준〉

탐색적 데이터 분석 (EDA)

▪ 수위(E지역)

- E 관측소의 수위 데이터
- 홍수사상별로 6개 데이터 집단의 수위(E지역)의 분포가 동일함을 확인

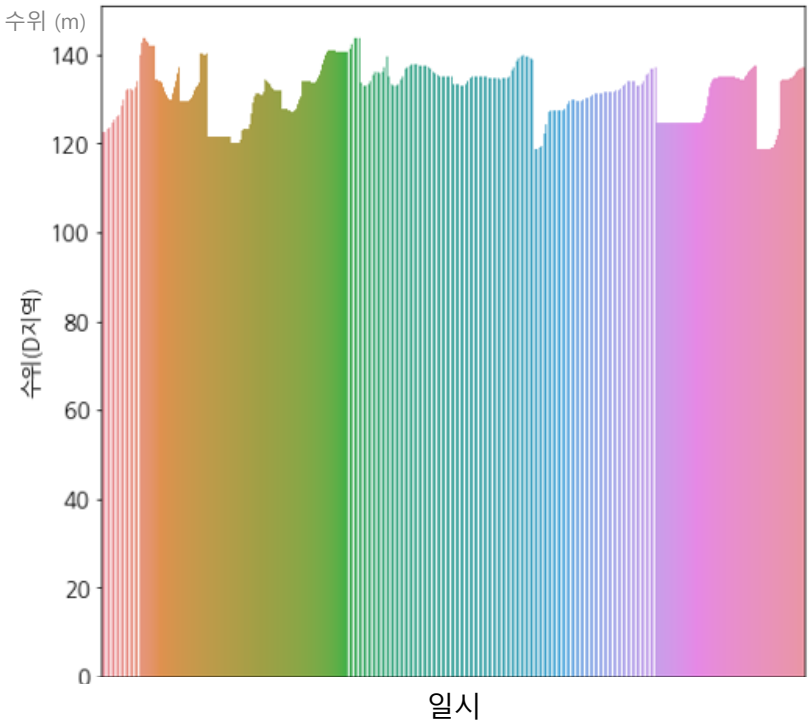


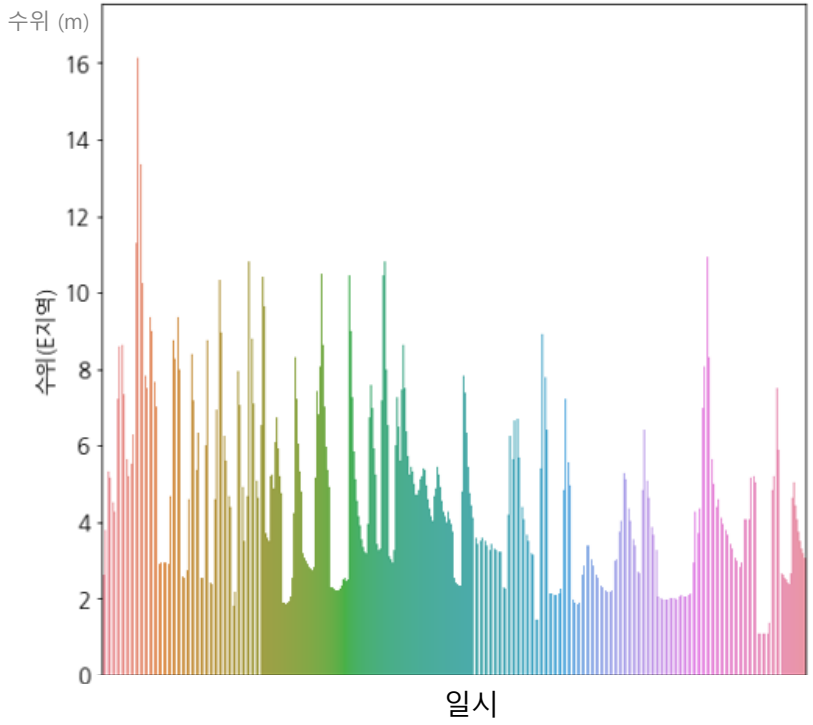
〈홍수사상번호 14번 기준〉

탐색적 데이터 분석 (EDA)

▪ 수위(D, E지역)

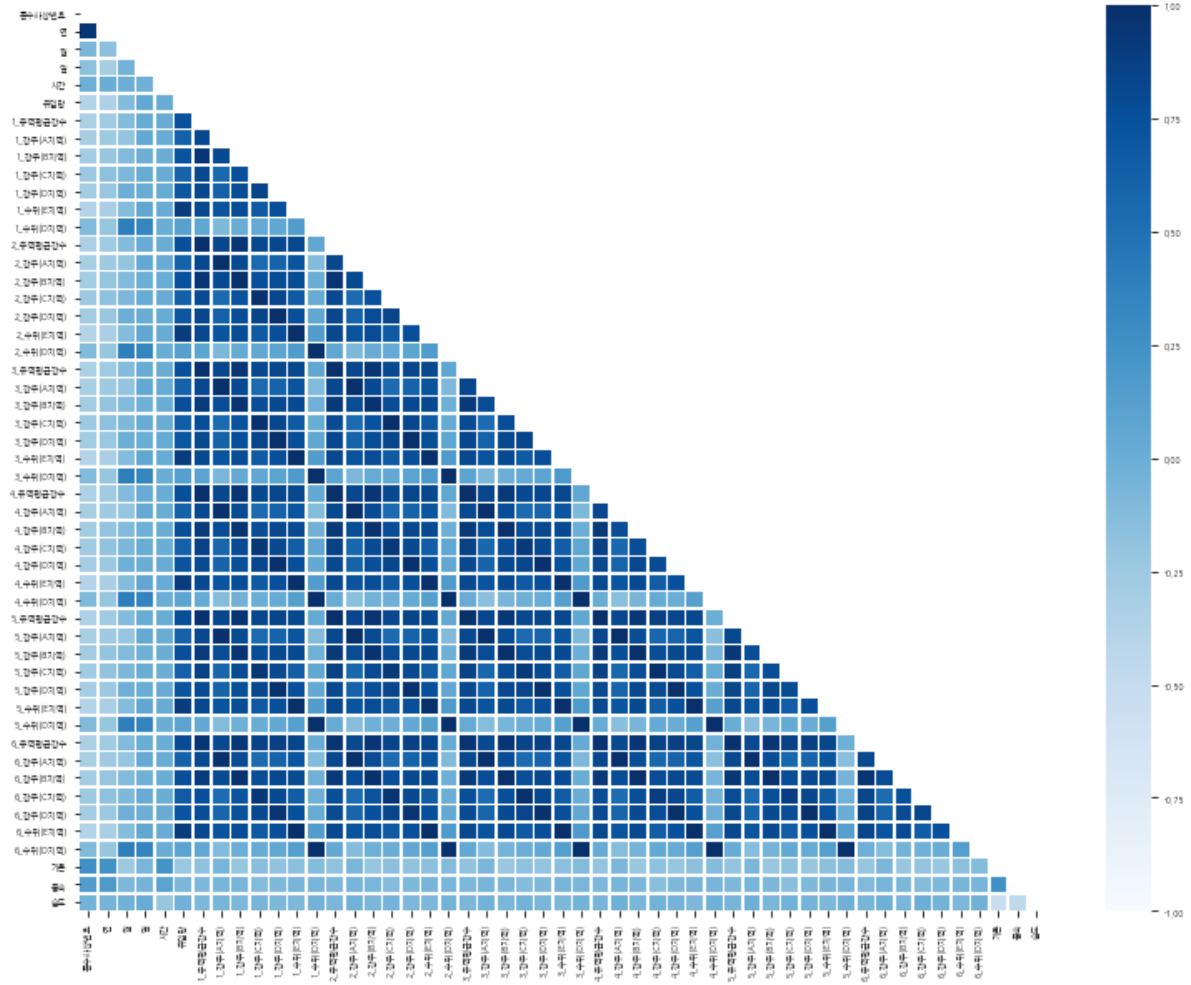
- 수위 D지역은 120 ~140(m)의 수위 값을 갖는 반면, 수위 E지역은 5~15(m)의 수위 값을 갖는다.
- 수위 D지역은 분포의 오르내림이 크지 않으며 일정한 수위 값을 가진다.





탐색적 데이터 분석 (EDA)

- **상관관계**
  - **제공된 데이터**와 **기온데이터**를 합친 데이터 프레임의 상관관계 확인

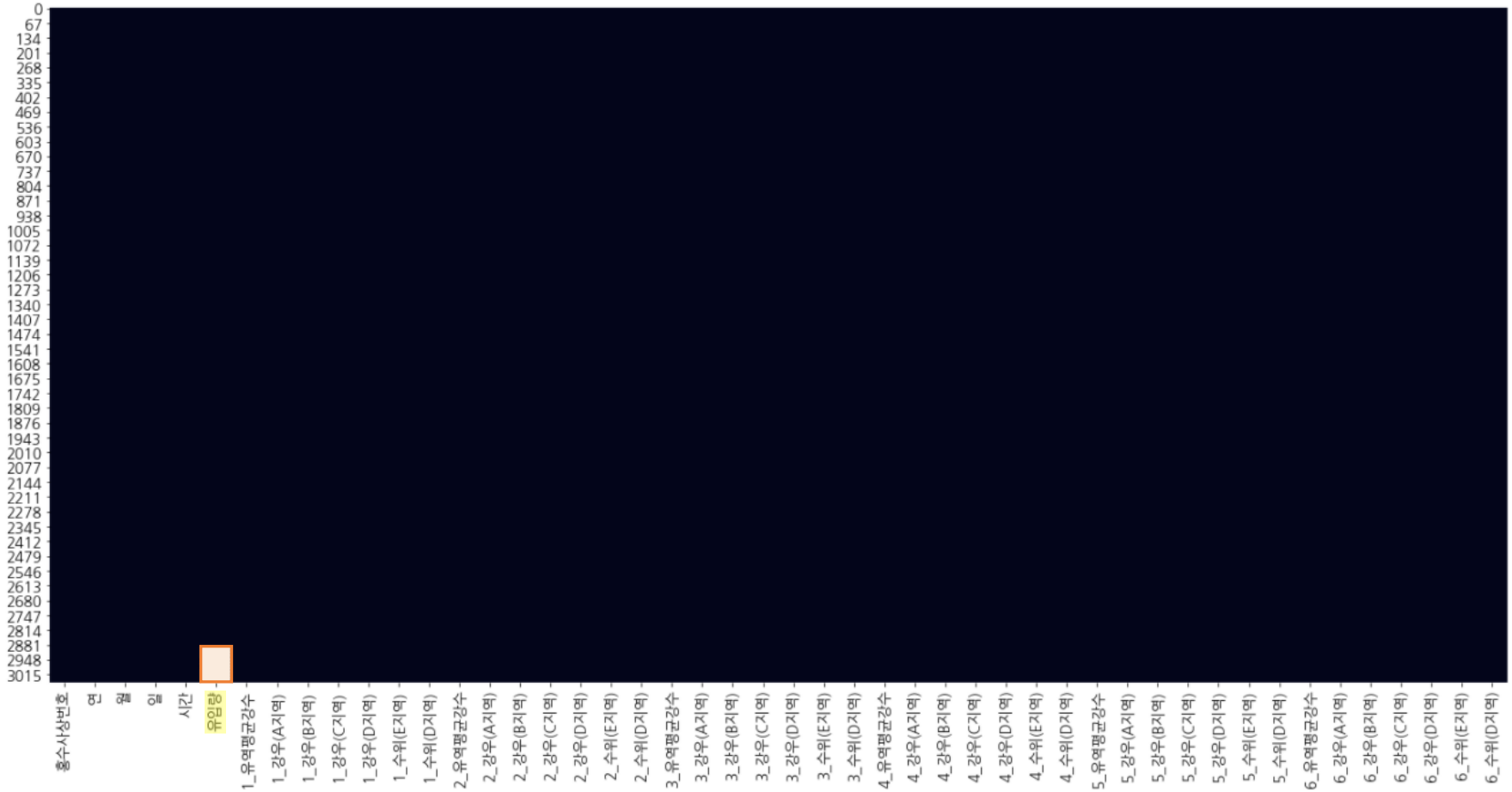


Ⅲ. 데이터 전처리

데이터 전처리

- **결측치**

  - target인 '홍수사상번호 26'의 유입량을 제외한 모든 구간 **결측치 없음**



데이터 전처리

- 기상데이터

	홍수사상번호	연	월	일	시간	일시
0	1	2006	7	10	8	2006-07-10 08:00
1	1	2006	7	10	9	2006-07-10 09:00
2	1	2006	7	10	10	2006-07-10 10:00
3	1	2006	7	10	11	2006-07-10 11:00
4	1	2006	7	10	12	2006-07-10 12:00
5	1	2006	7	10	13	2006-07-10 13:00
6	1	2006	7	10	14	2006-07-10 14:00
7	1	2006	7	10	15	2006-07-10 15:00
8	1	2006	7	10	16	2006-07-10 16:00
9	1	2006	7	10	17	2006-07-10 17:00
10	1	2006	7	10	18	2006-07-10 18:00
11	1	2006	7	10	19	2006-07-10 19:00
12	1	2006	7	10	20	2006-07-10 20:00
13	1	2006	7	10	21	2006-07-10 21:00
14	1	2006	7	10	22	2006-07-10 22:00
15	1	2006	7	10	23	2006-07-10 23:00
16	1	2006	7	10	24	2006-07-11 00:00
17	1	2006	7	11	1	2006-07-11 01:00
18	1	2006	7	11	2	2006-07-11 02:00
19	1	2006	7	11	3	2006-07-11 03:00
20	1	2006	7	11	4	2006-07-11 04:00
21	1	2006	7	11	5	2006-07-11 05:00

제공데이터의 연, 월, 일, 시를 기온데이터의 일시에 맞춤

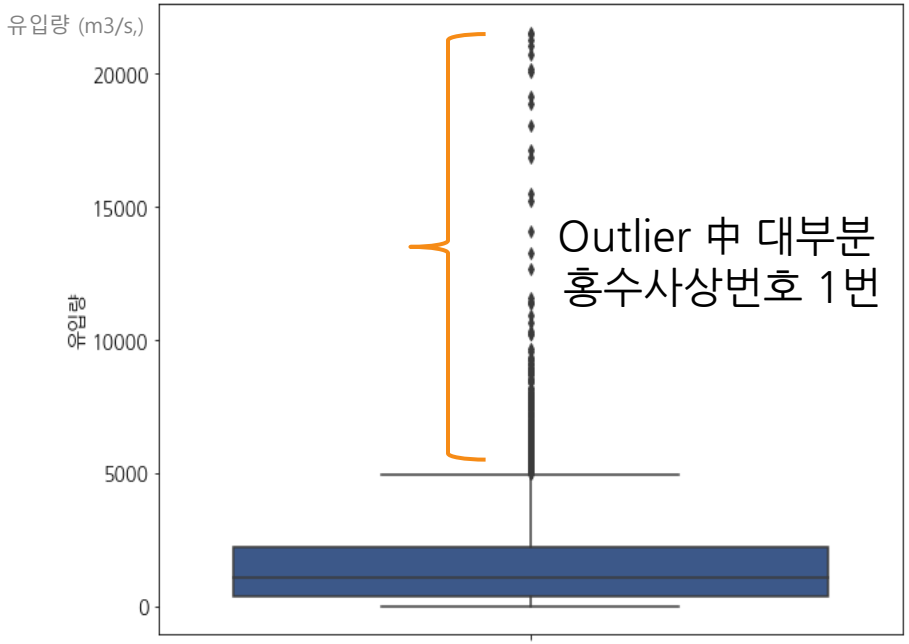
제공데이터의 시간 형식은 ‘**2006-07-10 24시**’
기온데이터는 ‘**2006-07-11 00시**’로 시간

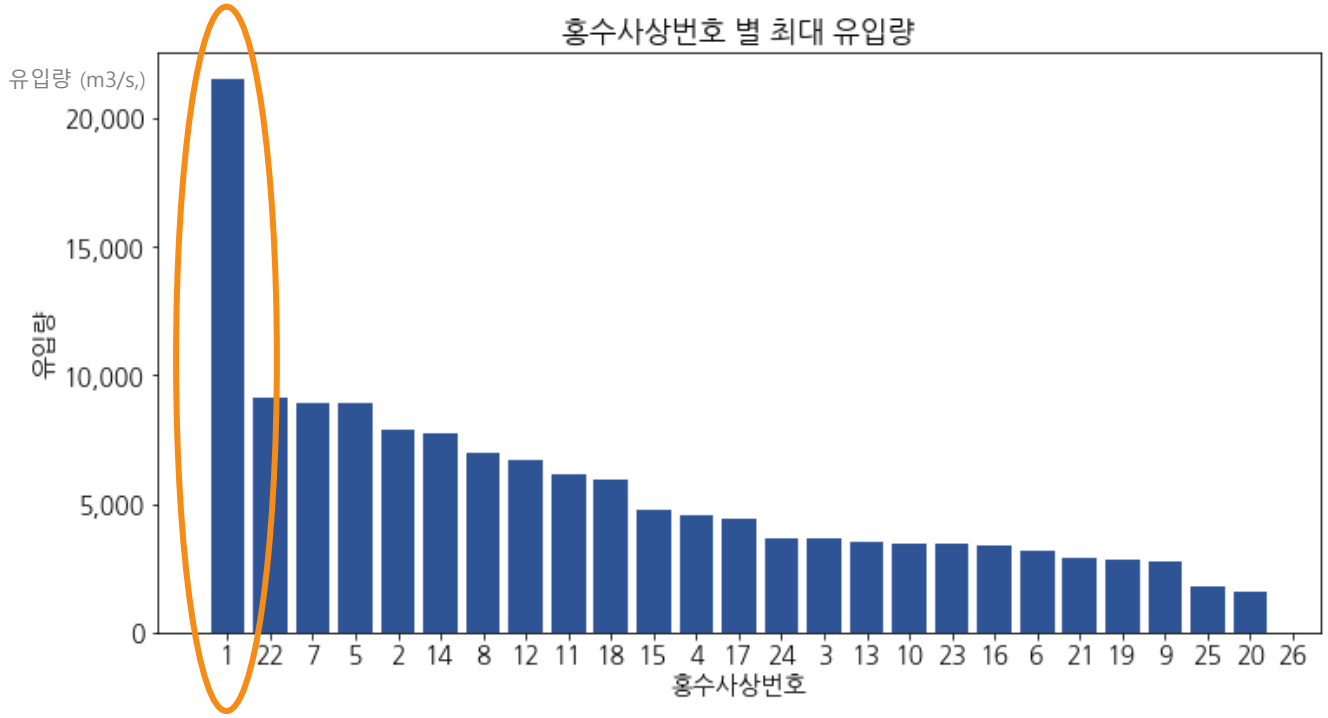
⇒ ‘일시’ 컬럼을 제거할 것이므로 형식을 변환하지 않음

데이터 전처리

- **이상치**
  - 유입량 데이터를 고려하였을 때, 홍수사상번호 1번의 유입량이 다른 유입량에 비해 현저히 높음
  - 또한, 유입량의 이상치 중 대부분이 홍수사상번호 1번에 속함

  ⇒ 따라서 홍수사상 1번을 이상치로 보고, 홍수사상번호 1번을 삭제





Ⅳ. Feature Engineering

Kmeans를 이용한 binning

- 데이터 그룹핑
  : K-means 알고리즘을 이용하여 유입량을 제외한 나머지 featur로 클러스터링 진행
- 각 클러스터와 거리 차이의 분산을 최소화 하는 방식이며,
  k는 실루엣 계수가 1에 가깝고 DB index는 값이 작은 2로 선정

```
n : 2
Silhouette Coefficient: 0.4044
Davies Bouldin Index: 1.1655
n : 3
Silhouette Coefficient: 0.2148
Davies Bouldin Index: 1.5682
n : 4
Silhouette Coefficient: 0.2239
Davies Bouldin Index: 1.3997
n : 5
Silhouette Coefficient: 0.2355
Davies Bouldin Index: 1.2887
n : 6
Silhouette Coefficient: 0.2509
Davies Bouldin Index: 1.2401
n : 7
Silhouette Coefficient: 0.2478
Davies Bouldin Index: 1.2990
n : 8
Silhouette Coefficient: 0.2474
Davies Bouldin Index: 1.3264
n : 9
Silhouette Coefficient: 0.2312
Davies Bouldin Index: 1.3481
```

K = 2 선정
(0 과 1 라벨링 진행)

	kmeans
0	0
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0

	kmeans
66	1
67	1
68	1
69	1
70	1
71	1
72	1
73	1
74	1
75	1
76	1

홍수 기간

- 홍수사상번호 별 홍수 지속 시간을 계산하여 홍수기간 컬럼을 생성
- 홍수사상 22번이 285시간으로 홍수지속 시간이 제일 길고, 홍수사상 4번이 34시간으로 가장 짧음

단위 (시간)

홍수사상번호	22	20	15	1	16	26	23	12	17	11	25	2	19	24	5	10	21	3	9	13	18	7	14	6	8	4
홍수기간	285	264	262	226	205	160	128	127	125	110	103	100	95	94	94	86	84	81	68	67	65	51	50	46	41	34

홍수기간 binning

- 앞서 생성한 홍수 기간 컬럼을 바탕으로 binning하여 또 다른 categorical 변수 생성
- 홍수기간에 대하여 8개의 구간을 설정하였다.

```
bins = [0,30, 50, 70, 100, 120, 160, 220]
data_feat['홍수_bin'] = np.digitize(data_feat['홍수기간'], bins)
data_feat.groupby('홍수_bin')[['홍수기간']].count().T
```

단위 (시간)

홍수_bin	1	2	3	4	5	6	7	8
홍수기간	754	494	416	480	196	263	287	161

▪ 절기

http://encykorea.aks.ac.kr/Contents/Item/E0049791

음력월	절기	양력일자	황경
5월(仲夏月)	망종(芒 種)	6월 6일경	75°
	하지(夏 至)	6월 21일경	90°
6월(季夏月)	소서(小 暑)	7월 7일경	105°
	대서(大 暑)	7월 23일경	120°
7월(孟秋月)	입추(立 秋)	8월 8일경	135°
	처서(處 暑)	8월 23일경	150°
8월(仲秋月)	백로(白 露)	9월 8일경	165°
	추분(秋 分)	9월 23일경	180°

	연	월	일	시간	월일	절기
0	2006	7	10	8	2	소서
1	2006	7	10	9	2	소서
2	2006	7	10	10	2	소서
3	2006	7	10	11	2	소서
4	2006	7	10	12	2	소서
...	...	...	...	...	...	...
3046	2018	7	7	17	-1	하지
3047	2018	7	7	18	-1	하지
3048	2018	7	7	19	-1	하지
3049	2018	7	7	20	-1	하지
3050	2018	7	7	21	-1	하지

연, 월, 일 대신 시간의 특성을 나타낼 수 있는 **'절기' 컬럼 추가**

그 중에서도 제공데이터의 기간에 해당하는 **하지, 소서, 대서, 입추, 처서, 백로, 추분**으로 분류

- **집단 7 생성**

  : 각각의 특징 변수들 간 최대값과 최소값을 뺀 나머지 4개의 관측치의 평균으로 구성

	1_유역평균강수	2_유역평균강수	3_유역평균강수	4_유역평균강수	5_유역평균강수	6_유역평균강수	7_유역평균강수
0	6.4000	6.3000	6.3000	6.4000	6.4000	6.4000	6.375000
1	6.3000	6.4000	6.4000	7.3000	7.3000	7.3000	6.850000
2	6.4000	7.3000	7.3000	8.2000	8.2000	8.2000	7.750000
3	7.3000	8.2000	8.2000	11.3000	11.3000	11.3000	9.750000
4	8.2000	11.3000	11.3000	14.4000	14.4000	14.4000	12.850000
...	...	...	...	...	...	...	...
3046	2.3689	2.3689	2.3689	2.3689	2.3689	2.1722	2.368900
3047	2.3689	2.3689	2.3689	2.3689	2.3689	2.0805	2.368900
3048	2.3689	2.3689	2.3689	2.3689	2.3689	2.0354	2.368900
3049	2.3689	2.3689	2.3689	2.3689	2.3488	1.8993	2.363875
3050	2.3689	2.3689	2.3689	2.3689	2.2615	1.8810	2.342050

3051 rows × 7 columns

- **집단 7 생성**

	7_유역평균강수	7_강우(A지역)	7_강우(B지역)	7_강우(C지역)	7_강우(D지역)	7_수위(D지역)
0	6.375000	7.0	7.0	7.50	8.00	122.597188
1	6.850000	7.0	8.0	8.50	9.00	122.592208
2	7.750000	7.0	9.0	8.75	9.50	122.587750
3	9.750000	8.0	10.0	12.00	11.50	122.586667
4	12.850000	10.5	12.0	14.00	13.25	122.582250
...	...	...	...	...	...	...
3046	2.368900	1.0	0.0	0.25	0.00	129.969104
3047	2.368900	1.0	0.0	0.25	0.00	129.982313
3048	2.368900	1.0	0.0	0.25	0.00	129.989375
3049	2.363875	1.0	0.0	0.25	0.00	129.996438
3050	2.342050	1.0	0.0	0.25	0.00	130.001937

3051 rows × 6 columns

▪ Rolling & Difference

: 데이터 집단 7에 대하여 각 관측치 별로 1시간 전, 2시간 전, 3시간 전의 유입량, 강우, 수위 추가

- Rolling

	7_유역평균강수_shift_1	7_강우(A지역)_shift_1	7_강우(B지역)_shift_1	7_강우(C지역)_shift_1	7_강우(D지역)_shift_1	7_수위(D지역)_shift_1	7_수위(E지역)_shift_1
0	5.900000	7.0	6.0	6.50	7.0	122.602167	2.55
1	6.375000	7.0	7.0	7.50	8.0	122.597188	2.54
2	6.850000	7.0	8.0	8.50	9.0	122.592208	2.53
3	7.750000	7.0	9.0	8.75	9.5	122.587750	2.53
4	9.750000	8.0	10.0	12.00	11.5	122.586667	2.53
...	...	...	...	...	...	...	...
3046	2.368900	1.0	0.0	0.50	0.0	129.957229	3.18
3047	2.368900	1.0	0.0	0.25	0.0	129.969104	3.16
3048	2.368900	1.0	0.0	0.25	0.0	129.982313	3.15
3049	2.368900	1.0	0.0	0.25	0.0	129.989375	3.13
3050	2.363875	1.0	0.0	0.25	0.0	129.996438	3.11

3051 rows × 28 columns

- Difference

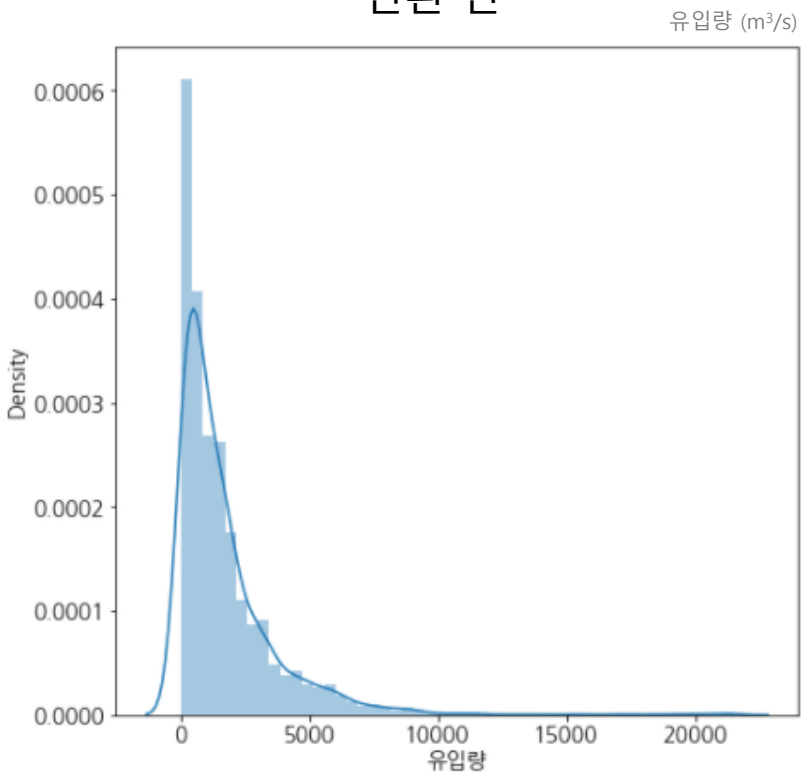
7_유역평균강수_diff	7_강우(A지역)_diff	7_강우(B지역)_diff	7_강우(C지역)_diff	7_강우(D지역)_diff	7_수위(D지역)_diff	7_수위(E지역)_diff
0.475000	0.0	1.0	1.00	1.00	-0.004979	-0.01
0.475000	0.0	1.0	1.00	1.00	-0.004979	-0.01
0.900000	0.0	1.0	0.25	0.50	-0.004458	0.00
2.000000	1.0	1.0	3.25	2.00	-0.001083	0.00
3.100000	2.5	2.0	2.00	1.75	-0.004417	0.00
...	...	...	...	...	...	...
0.000000	0.0	0.0	-0.25	0.00	0.011875	-0.02
0.000000	0.0	0.0	0.00	0.00	0.013208	-0.01
0.000000	0.0	0.0	0.00	0.00	0.007062	-0.02
-0.005025	0.0	0.0	0.00	0.00	0.007063	-0.02
-0.021825	0.0	0.0	0.00	0.00	0.005500	-0.01

불완전한 시계열 특성을 띄는 제공데이터

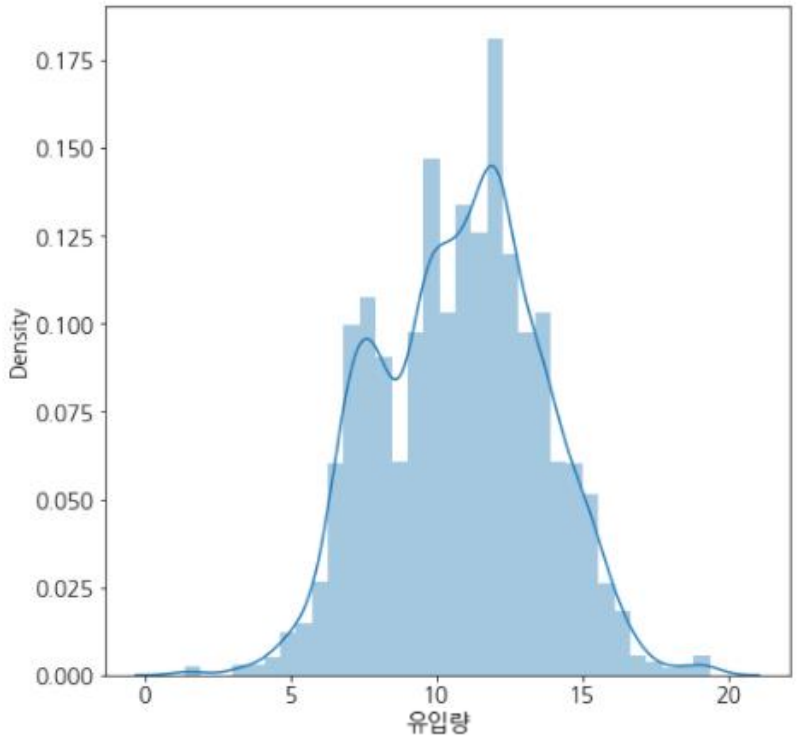
⇨ rolling과 difference를 통해 **과거의 정보를 추가적으로 사용**하여 예측도 향상 도모

- box_cox 변환 (유입량)

변환 전



변환 후

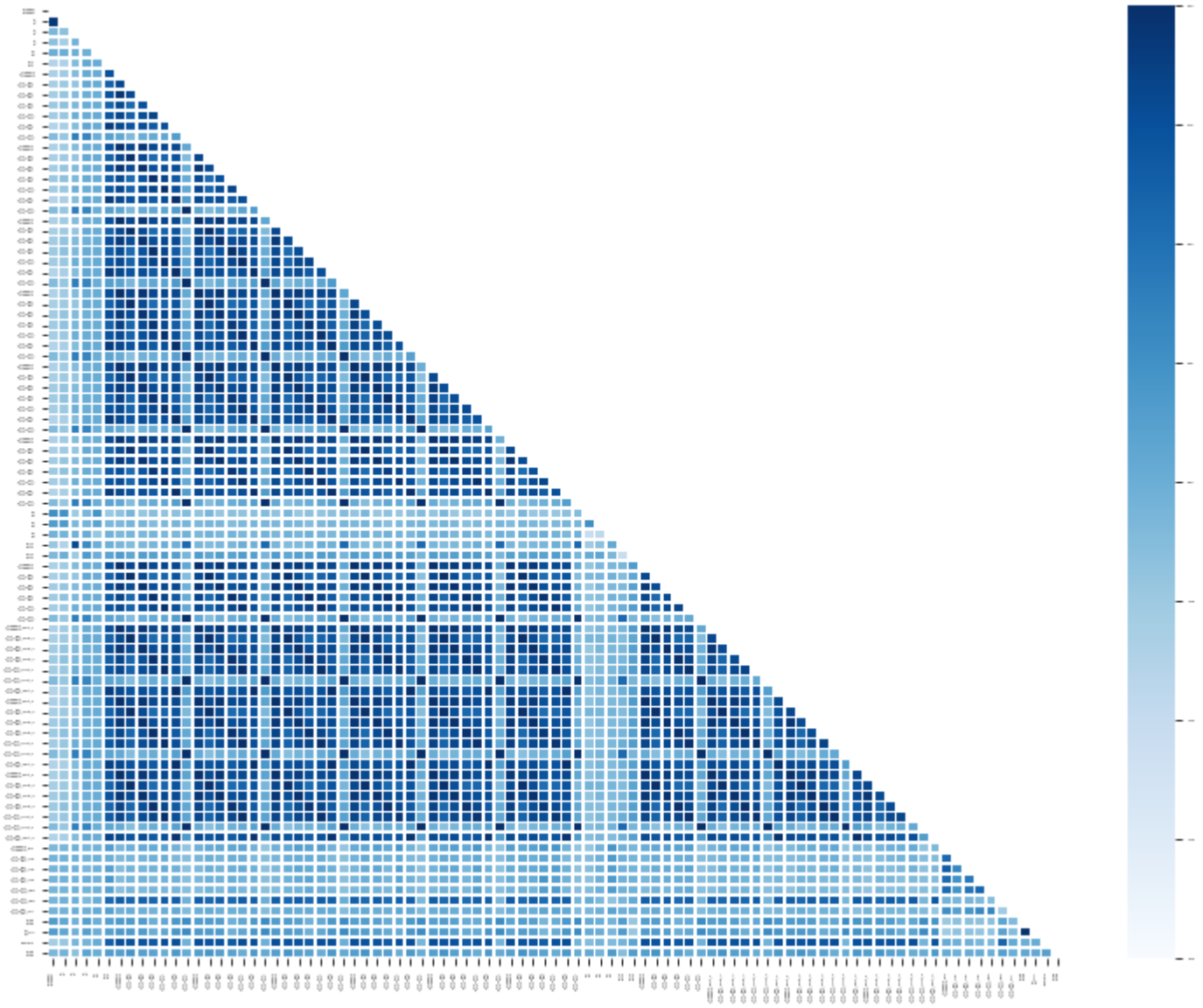


Box-Cox 변환은 정규 분포와 매우 유사하도록 데이터를 변환

⇨ **백색 잡음 제거**를 통해 모델의 예측력 ↑

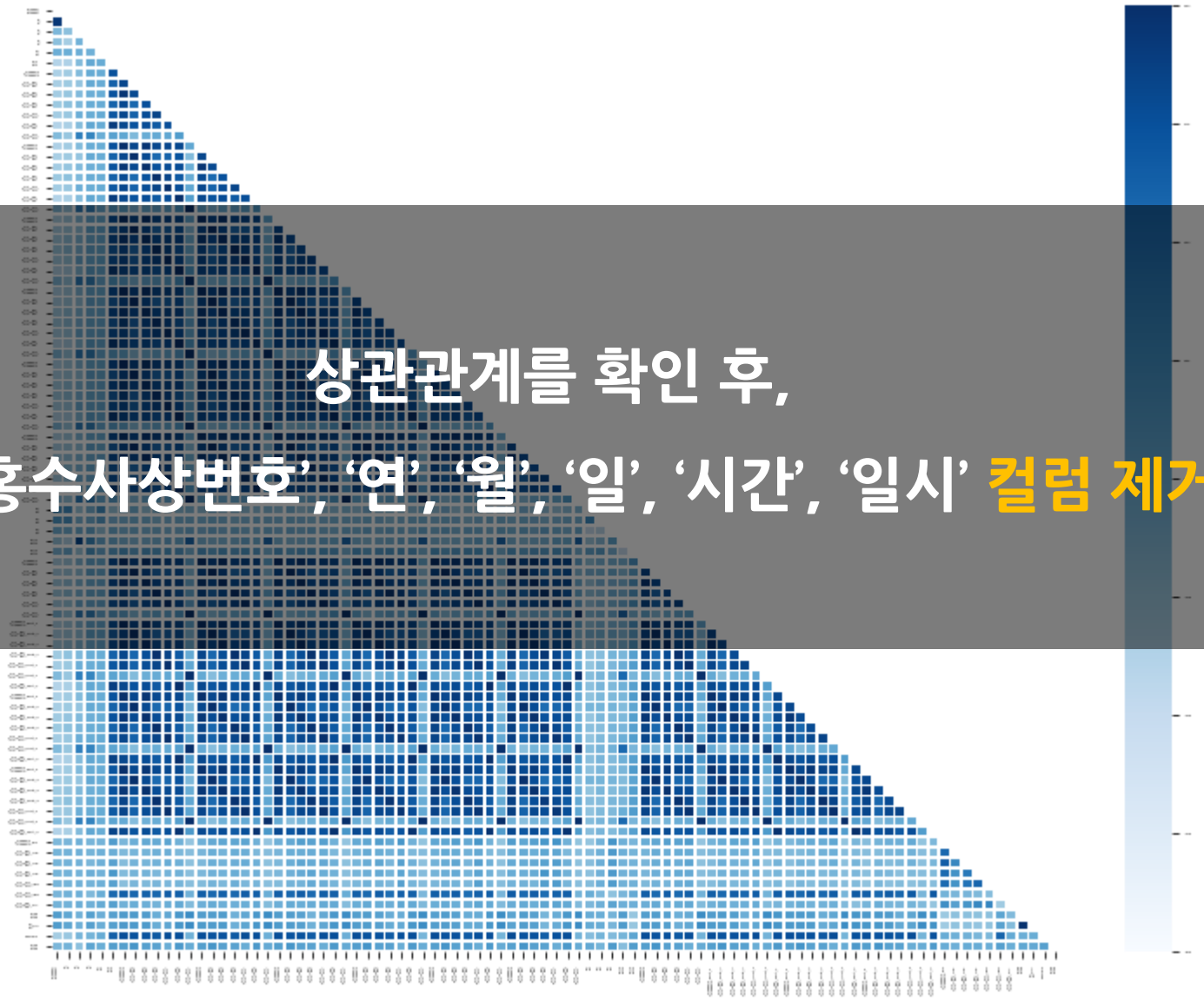
Feature Engineering

- Feature Selection



Feature Engineering

- Feature Selection



상관관계를 확인 후,

⇒ ‘홍수사상번호’, ‘연’, ‘월’, ‘일’, ‘시간’, ‘일시’ **컬럼 제거**

V. Modeling

Data Scaling

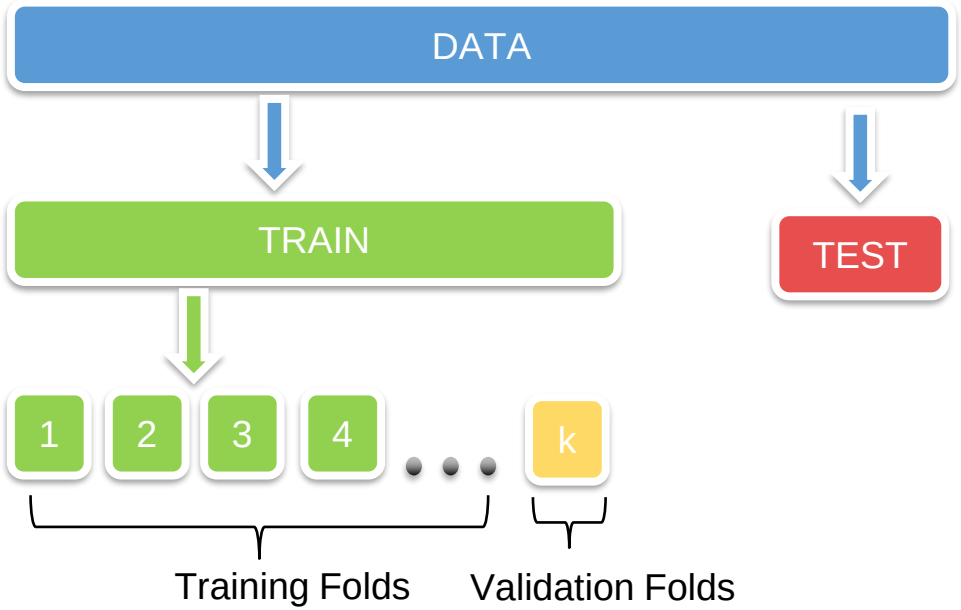
- StandardScaler를 이용하여 수치형 칼럼들에 대하여 스케일 적용
- 타겟 변수인 유입량의 경우 Box-Cox 변환을 진행

1_유역평균강수	1_강우(A지역)	1_강우(B지역)	1_강우(C지역)	1_강우(D지역)	1_수위(E지역)	1_수위(D지역)	2_유역평균강수	2_강우(A지역)	2_강우(B지역)	2_강우(C지역)
6.4000	7	7	7	8	2.54	122.56875	6.3000	7	7	7
6.3000	7	8	7	8	2.53	122.56250	6.4000	7	8	7
6.4000	7	9	7	8	2.53	122.55625	7.3000	7	9	7
7.3000	7	10	7	8	2.53	122.55625	8.2000	7	10	8
8.2000	7	12	8	10	2.53	122.55625	11.3000	9	12	10
...	...	...	...	...	...	...	...	...	...	...
2.3689	1	0	0	0	3.16	129.99375	2.3689	1	0	0
2.3689	1	0	0	0	3.15	130.00625	2.3689	1	0	0
2.3689	1	0	0	0	3.13	130.01250	2.3689	1	0	0
2.3689	1	0	0	0	3.11	130.01875	2.3689	1	0	0
2.3689	1	0	0	0	3.10	130.01875	2.3689	1	0	0

1_유역평균강수	1_강우(A지역)	1_강우(B지역)	1_강우(C지역)	1_강우(D지역)	1_수위(E지역)	1_수위(D지역)	2_유역평균강수	2_강우(A지역)	2_강우(B지역)	2_강우(C지역)
0.019488	0.021277	0.019074	0.027668	0.030534	0.093930	0.153560	0.019184	0.020772	0.019074	0.027888
0.019184	0.021277	0.021798	0.027668	0.030534	0.093291	0.153312	0.019488	0.020772	0.021798	0.027888
0.019488	0.021277	0.024523	0.027668	0.030534	0.093291	0.153064	0.022229	0.020772	0.024523	0.027888
0.022229	0.021277	0.027248	0.027668	0.030534	0.093291	0.153064	0.024970	0.020772	0.027248	0.031873
0.024970	0.021277	0.032698	0.031621	0.038168	0.093291	0.153064	0.034409	0.026706	0.032698	0.039841
...	...	...	...	...	...	...	...	...	...	...
0.069378	0.018237	0.000000	0.003953	0.003817	0.125240	0.733069	0.025148	0.017804	0.000000	0.003984
0.025148	0.006079	0.000000	0.003953	0.003817	0.123323	0.733565	0.012512	0.005935	0.000000	0.003984
0.012512	0.003040	0.000000	0.003953	0.003817	0.122045	0.733813	0.010309	0.002967	0.000000	0.003984
0.010309	0.003040	0.000000	0.003953	0.003817	0.120767	0.734061	0.010000	0.002967	0.000000	0.003984
0.010000	0.003040	0.000000	0.003953	0.003817	0.119489	0.734557	0.008378	0.002967	0.000000	0.003984

Cross Validation

- 데이터의 크기가 크지 않고, test set에 과적합을 방지하고자 교차검증으로 모델링을 진행
- 홍수사상 별로 홍수 사건이 나뉘어지며, 연속적으로 데이터가 측정되었기 때문에 랜덤하게 fold를 나누지 않고, 홍수사상번호 별로 fold를 나누어 교차검증을 진행하였다.



Bayesian Optimization

- 하이퍼 파라미터 튜닝을 위해 베이지안 최적화를 이용하여 모델링 진행

Machine Learning Model 선택한 이유

- 설명가능한 AI(XAI)를 통한 모델 결과의 효과적인 이해
- 시간의 흐름을 반영하기 위해 차분과 롤링을 사용
- 사용한 ML 모델
  : Linear Regressor, RandomForest, XGB, Extra Tree, LGBM

Deep Learning을 선택한 이유

- 홍수사상별 일정한 시간동안 수집된 일련의 순차적 데이터 셋이라고 판단

- 시계열 분석을 통한 예측 모형 구축하기 위함
- 사용한 딥러닝 모델
  : LSTM, GRU

▪ Hyperparameter Tunning

- LGBM Gradient Boosting

```
lgb.LGBMRegressor(random_state = 42,
          n_jobs = -1,
          feature_fraction = 0.4333407057741526,
          learning_rate = 0.19063571821788408,
          max_depth = 6,
          min_child_weight = 1.7959754525911098,
          n_estimators = 5780,
          num_leaves = 131,
          reg_alpha = 0.05750277604651747,
          reg_lambda = 0.8575143843171859,
          subsample = 0.8005575058716043
          )
```

- LGBM Gradient-Based One-side Sampling

```
lgb.LGBMRegressor(random_state = 42,
          n_jobs = -1,
          boosting_type = 'goss',
          feature_fraction = 0.7770000834225976,
          learning_rate = 0.14816927376596317,
          max_depth = 6,
          min_child_weight = 0.22634193152740667,
          n_estimators = 6414,
          num_leaves = 186,
          reg_alpha = 0.023192578393987293,
          reg_lambda = 0.9575539951600276,
          subsample = 0.6051769440265253
          )
```

- XGBoost1

```
xgb.XGBRegressor(random_state = 42,
                 n_jobs = -1,
                 eval_metric = 'rmse',
                 objective='reg:squarederror',
                 colsample_bytree = 0.38637607534866675,
                 gamma = 67.98849651049338,
                 learning_rate = 0.014866598172025397,
                 max_depth = 5,
                 min_child_weight = 1.7224699992946557,
                 n_estimators = 9896,
                 subsample = 0.726566924360085
                  )
```

- XGBoost2 : reg추가

```
xgb.XGBRegressor(colsample_bytree = 0.99,
                         gamma = 0.0,
                         learning_rate = 0.05,
                         max_depth = 10,
                         min_child_weight = 0.9215786233997231,
                         n_estimators = 6038,
                         reg_alpha = 0.36516267277203135,
                         reg_lambda = 0.01,
                         subsample = 0.7611671326691636)
```

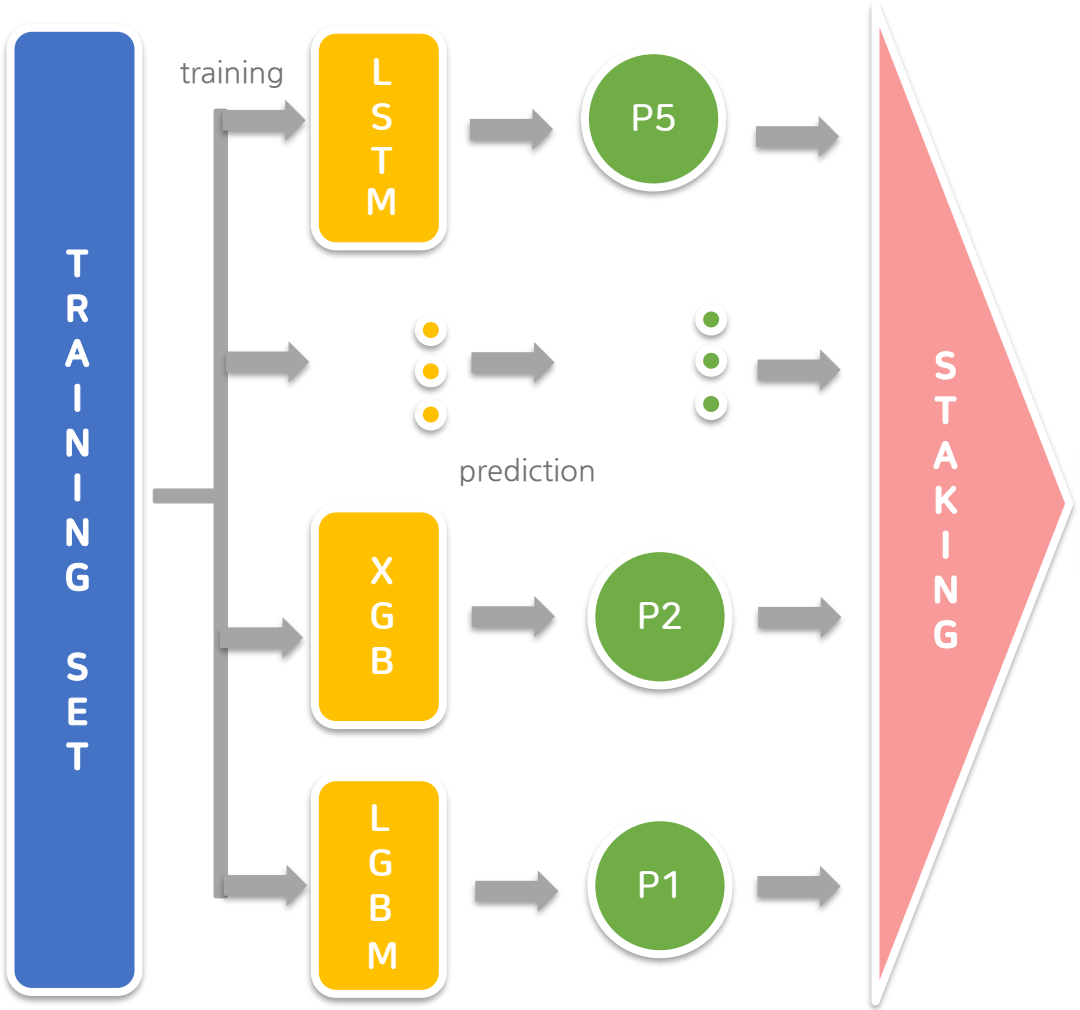
- **Hyperparameter Tunning**

- LSTM

```
window_size = 5
model.add(LSTM(526, activation = 'tanh',
               input_shape = x_train[0].shape))
model.add(Dense(256))
model.add(Dense(1, activation='linear'))
model.summary()

model.compile(loss='mse', optimizer='adam', metrics=['mae'])
early_stop = EarlyStopping(monitor = 'val_loss', patience = 5)
model.fit(x_train, y_train, validation_data = (x_valid, y_valid), epochs = 1000, verbose=0,
          batch_size = 200, callbacks = [early_stop])
```

- STACKING

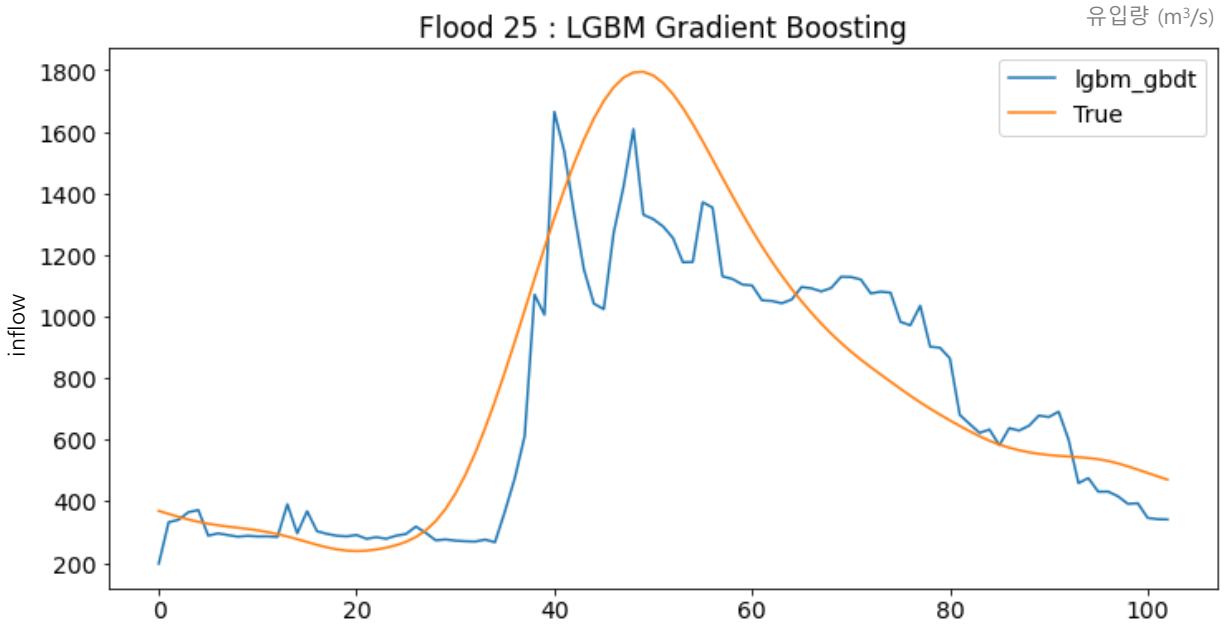


⇒ LGBM_goss, LGBM_gbdt, XGBoost1, XGBoost2, LSTM의 predict 값을 이용한 stacking

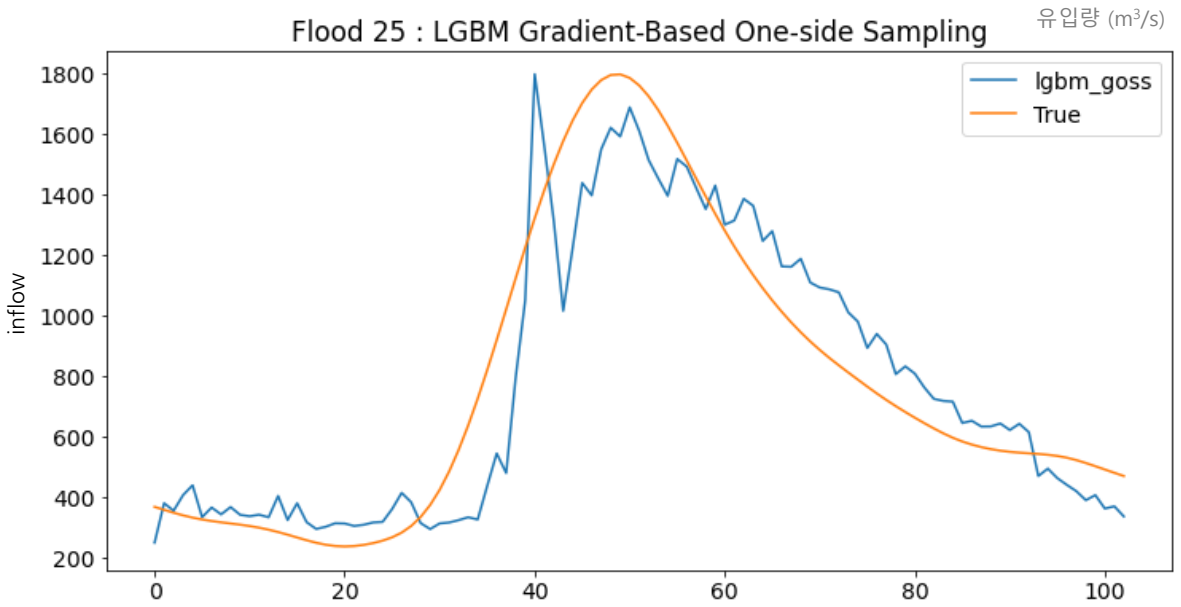
VI. 성능 평가

홍수사상번호 25번 예측 그래프

- LGBM Gradient Boosting



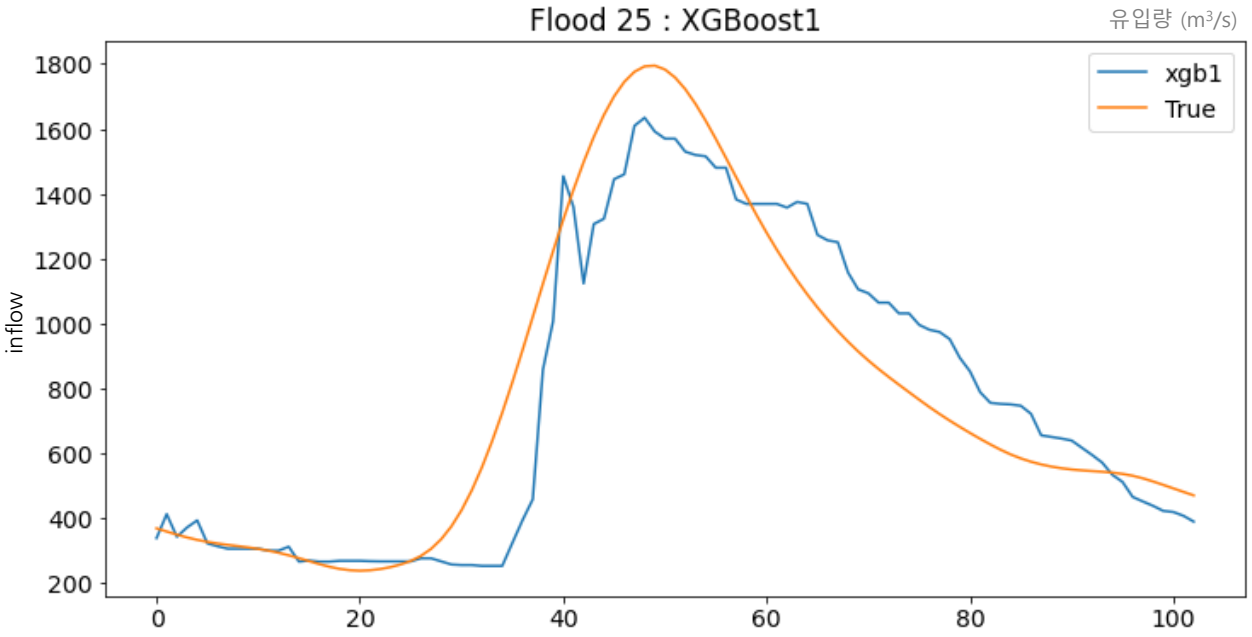
- LGBM Gradient-Based One-side Sampling



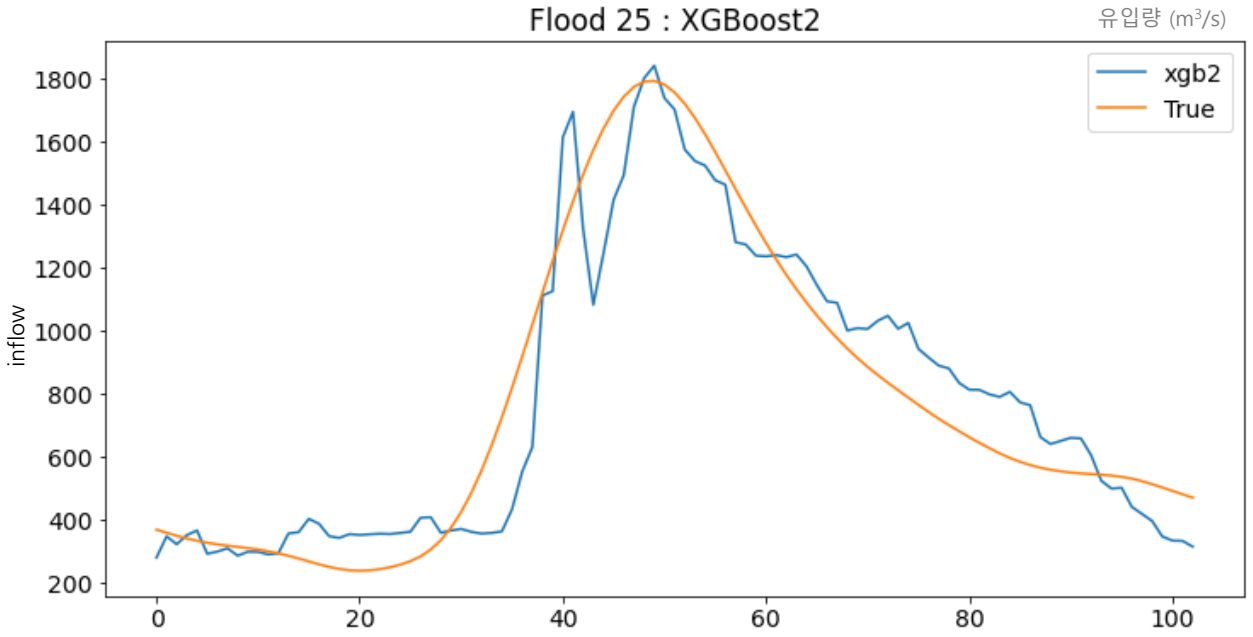
성능 평가

홍수사상번호 25번 예측 그래프

- XGBoost1



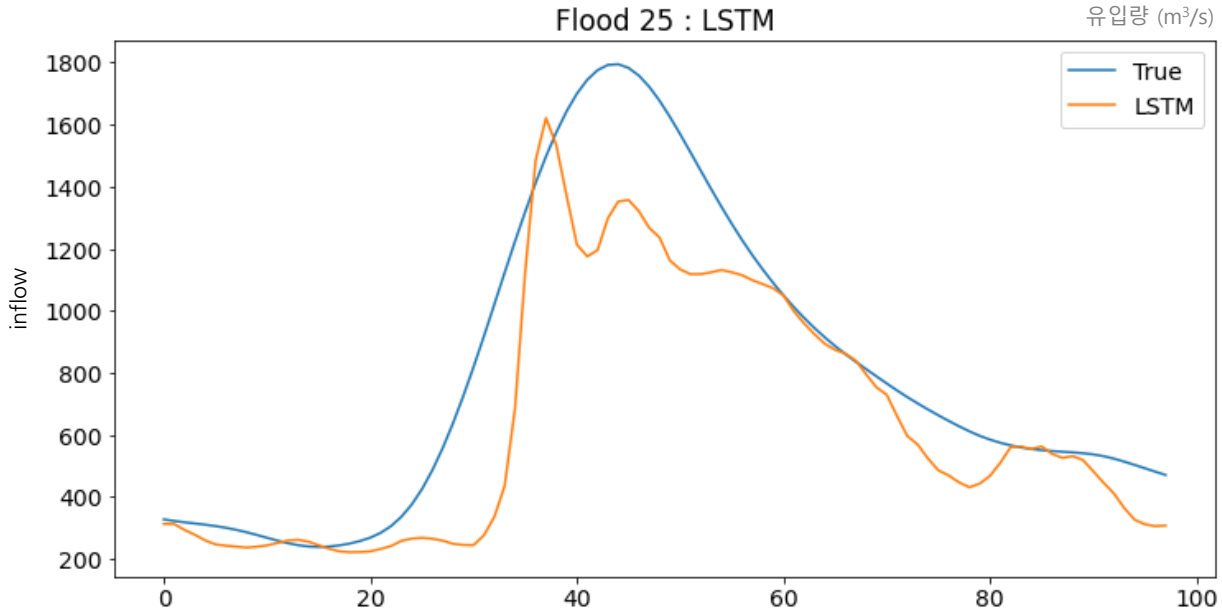
- XGBoost2 : reg추가



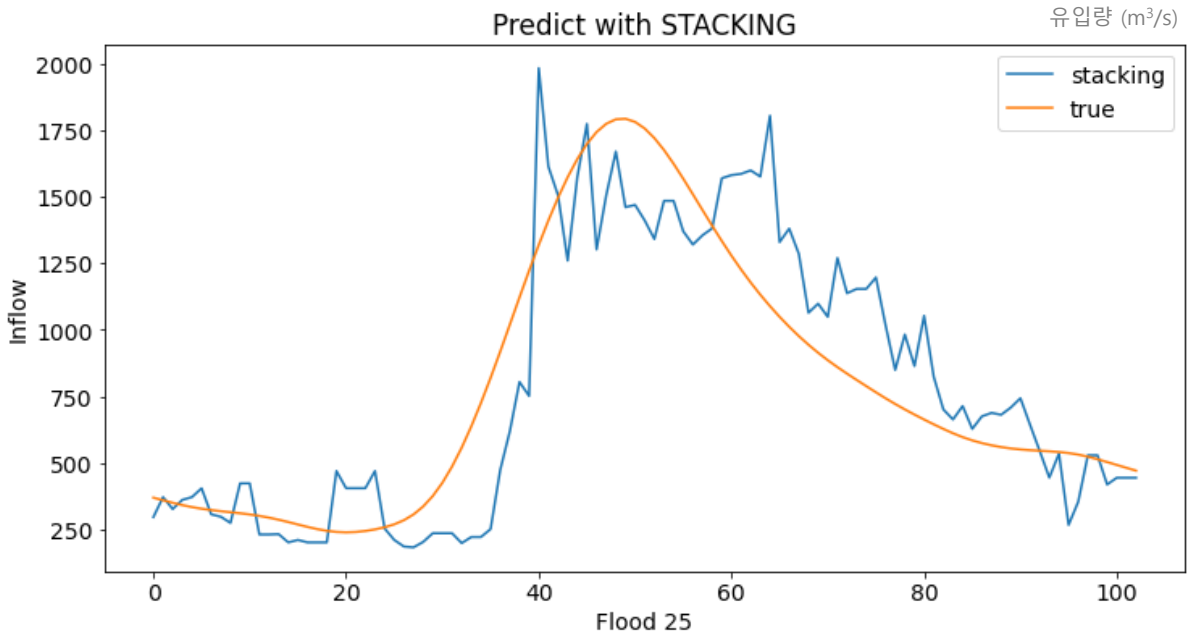
성능 평가

홍수사상번호 25번 예측 그래프

- LSTM

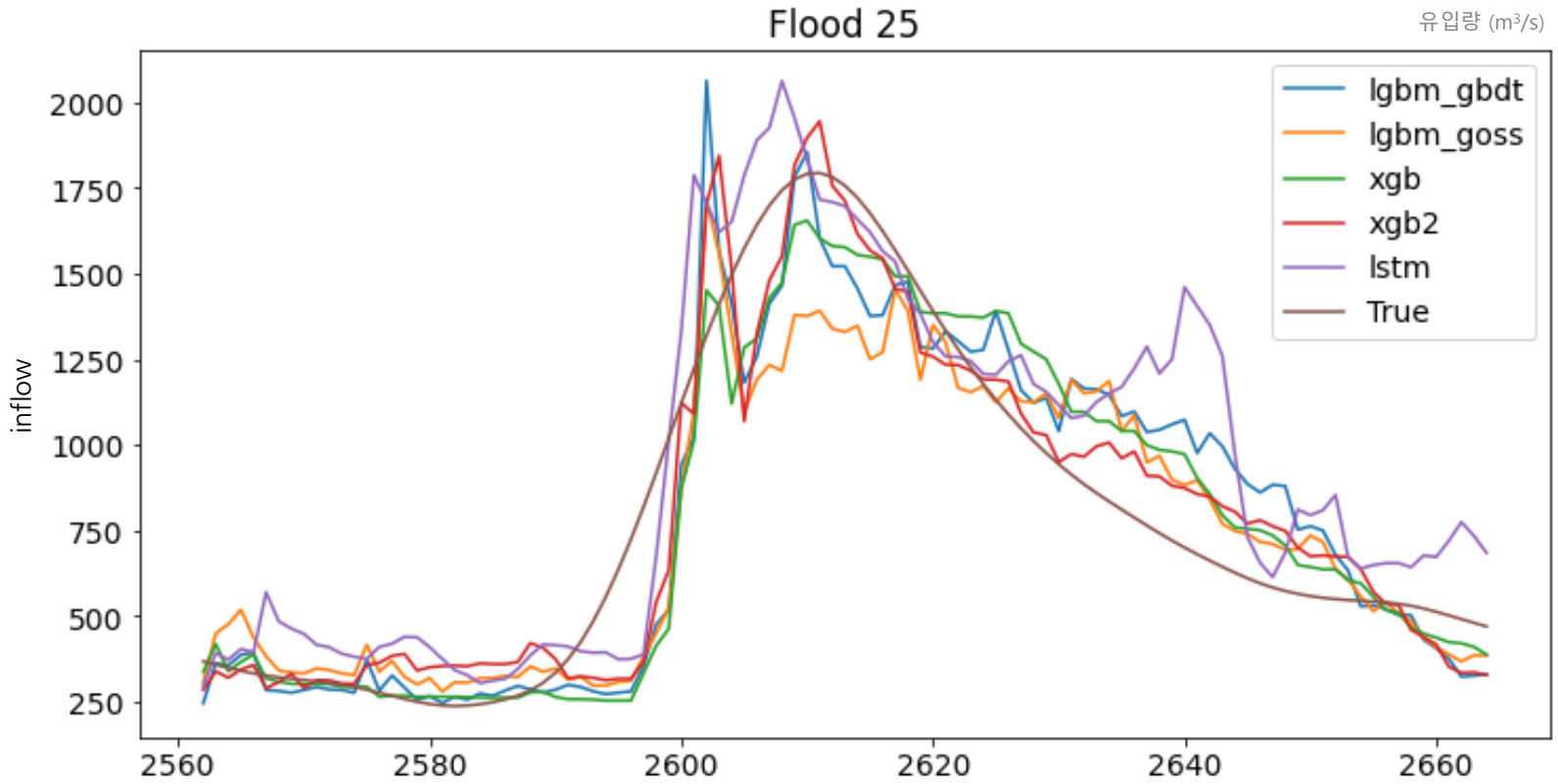


- STACKING



성능 평가

- **모델의 예측 그래프**
  - 홍수사상번호 25번



성능 평가

- **METRIC**

① RMSE (Root Mean Square Error)

- 예측값과 실제값 사이의 잔차 값을 기준으로 모델의 정확도를 측정하는 지표
- 예측하려는 값의 크기에 따라 에러가 의존적인 경향 존재

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

② MAPE (Mean Absolute Percentage Error)

- 실제값 대비 예측값과 실제값 사이의 잔차 비율 평균
- 홍수사상별로 유입량이 다르기에, RMSE의 크기 의존적인 에러의 단점을 보완하고자 사용

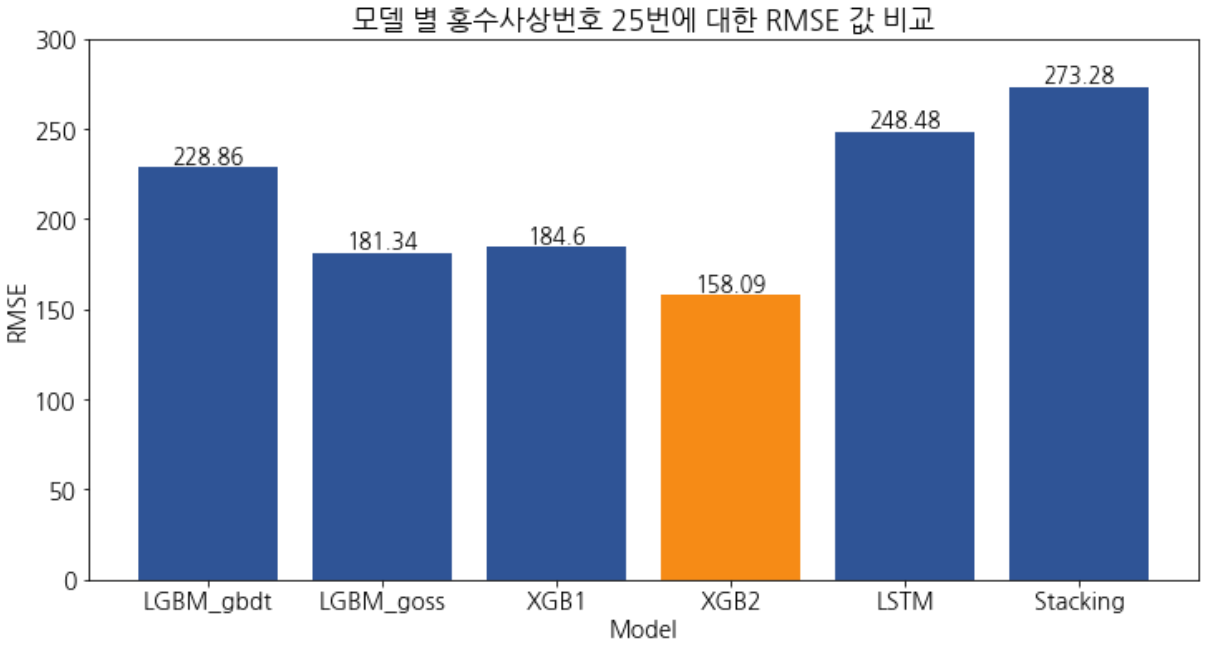
$$\mathrm{M} = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

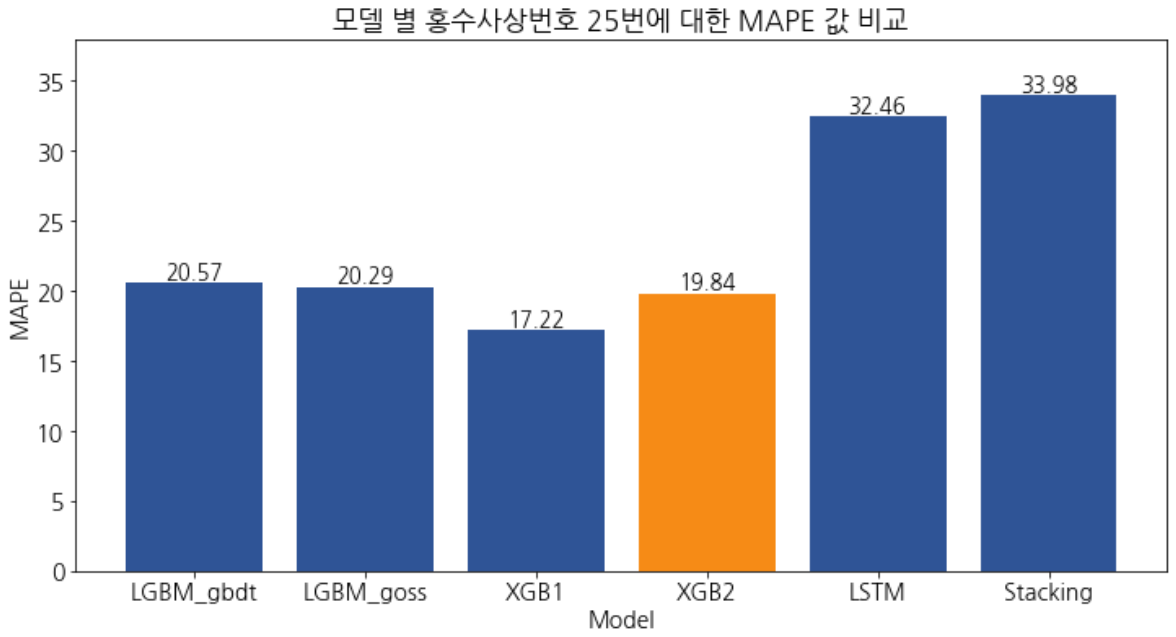
A_t: 실제값, F_t : 예측값

성능 평가

▪ 성능 평가

- 교차검증 진행 후, 홍수사상번호 25번에 대하여 RMSE값과 MAPE값을 비교
- **XGBoost2 모델**의 **RMSE값이 158.09로 최소, MAPE값은 19.84**로 두번째로 작음

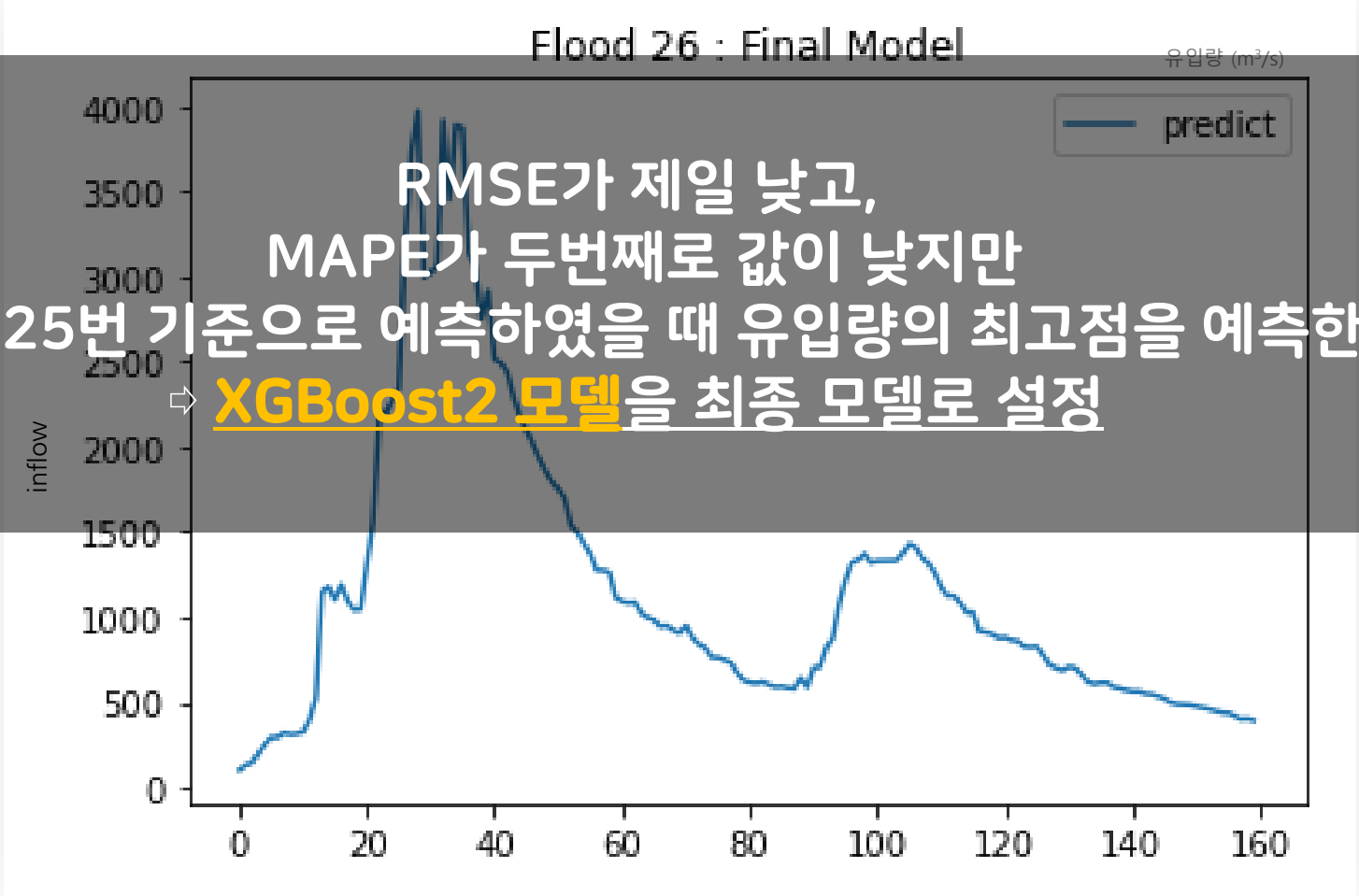




Ⅶ. 결론 및 토론

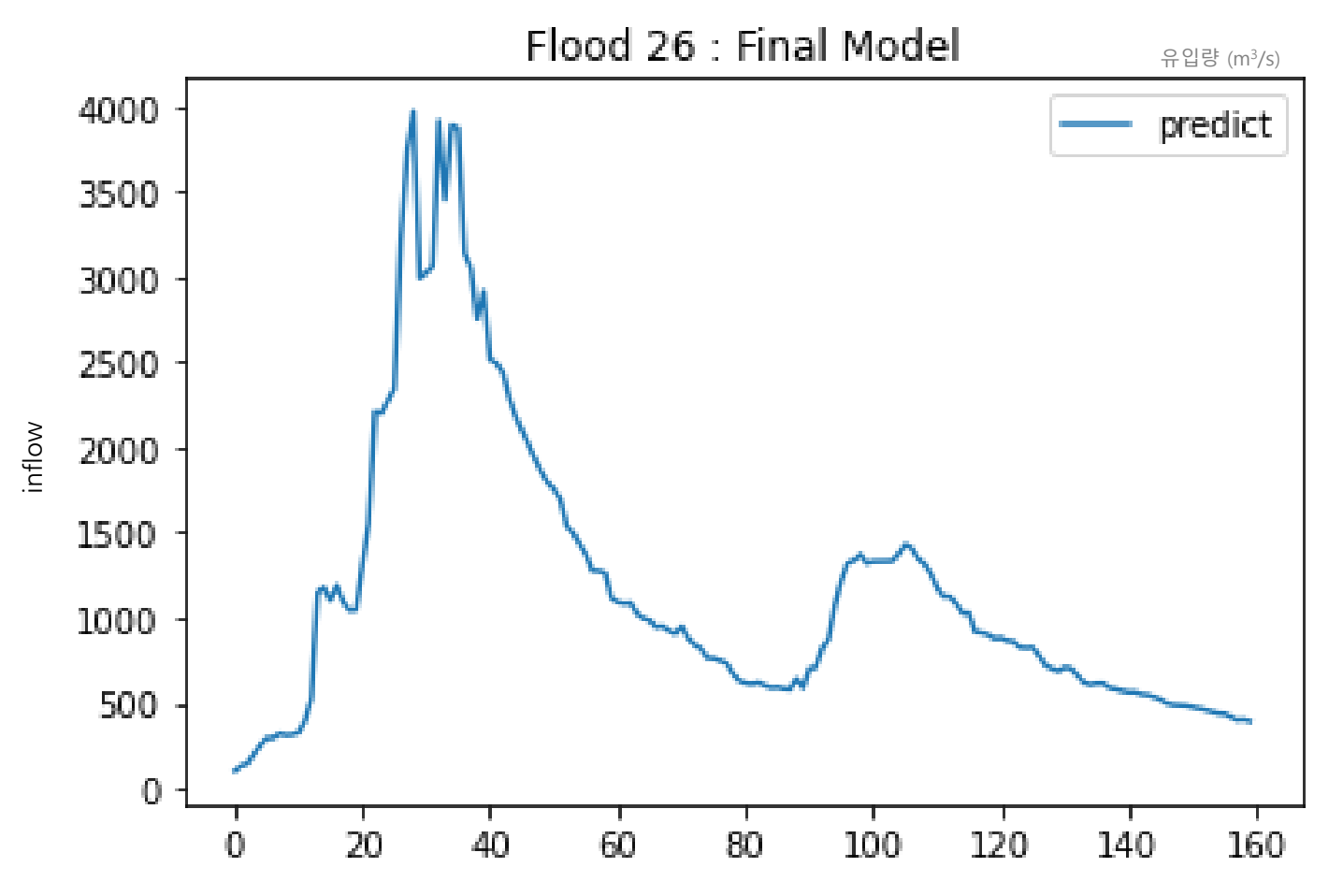
결론 및 토론

- **최종 모델 선택**
  - 홍수사상번호 26번을 XGBoost2로 예측한 그래프



RMSE가 제일 낮고,
MAPE가 두번째로 값이 낮지만
홍수사상번호 25번 기준으로 예측하였을 때 유입량의 최고점을 예측한 점을 고려
⇨ **XGBoost2 모델**을 최종 모델로 설정

- **최종 모델 선택**
  - 홍수사상번호 26번을 XGBoost2로 예측한 그래프

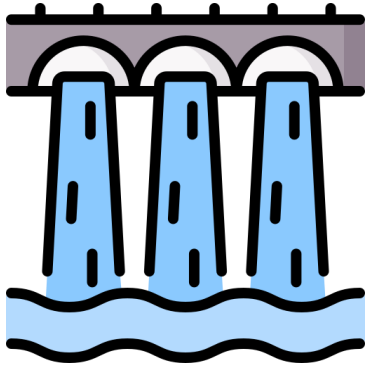


결론 및 토론

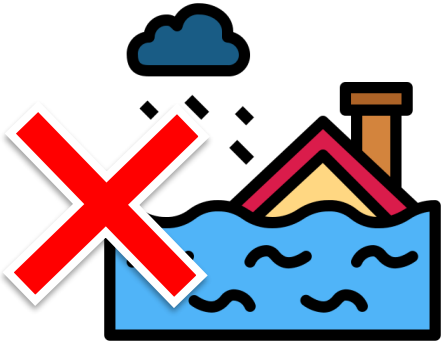
- 기대효과

COSFIM 모델

XGB 모델



효과적인 댐 운영 관리



홍수 피해 최소화

한계점

- 주어진 데이터에서 지역적인 특성이 배제되어 기온 데이터 뿐만 아니라 더 많은 외부 데이터 활용의 한계
- 제공된 데이터는 홍수사상 기준으로 이뤄진 데이터로, 홍수가 발생하지 않는 날들의 정보가 없어 보다 나은 예측의 한계
- 데이터의 관측 시간 단위가 '시간'으로 다소 크기 때문에 데이터의 볼륨이 작다는 한계

REFERENCE

https://keras.io/

https://scikit-learn.org

https://pycaret.org/

https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

https://www.water.or.kr/disaster/general/flood/flood03_datail06.do

https://www.youtube.com/watch?v=oNLaw2Q8Irw&list=PL9mhQYIlKEhd60Qq4r2yC7xYKIhs97FfC

https://www.youtube.com/watch?v=hCl8zTYM4So&list=PLSN_PltQeOyjnE4AnJyQUlHXNwE_hVtKL

https://wooono.tistory.com/102

https://www.weather.go.kr/w/index.do

https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMRegressor.html

https://pandas.pydata.org/docs/index.html

https://stackoverflow.com/questions/45361559/feature-importance-chart-in-neural-network-using-keras-in-python

https://ko.wikipedia.org/wiki/%EC%A0%88%EA%B8%B0

https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/

감사합니다