Chapter. 08

## 시작에 앞서: 탐색적 데이터 분석이란?

# 정의 및 중요성

FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

# I 정의

- 탐색적 데이터 분석 (Exploratory Data Analysis; EDA)란 그래프를 통한 **시각화**와 **통계 분석** 등을 바탕으로 수집한 데이터를 다양한 각도에서 **관찰**하고 **이해**하는 방법

# I 필요성 및 효과

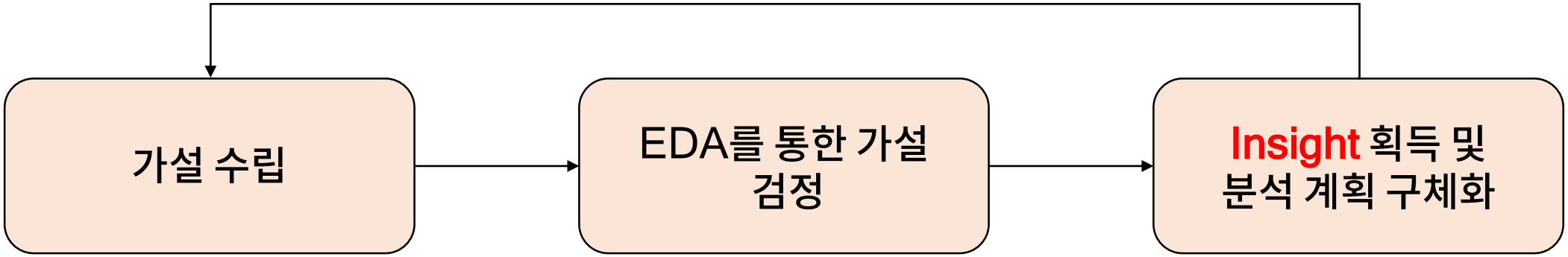
- 데이터를 여러 각도에서 살펴보면서 데이터의 전체적인 양상과 보이지 않던 현상을 더 잘 **이해**할 수 있도록 도와줌

- 문제 정의 단계에서 발견하지 못한 패턴을 발견하고, 이를 바탕으로 데이터 전처리 및 모델에 관한 **가설을 추가하거나 수정**할 수 있음

- 즉, 본격적인 데이터 분석에 앞서 구체적인 분석 계획을 수립하는데 도움이 됨

# 가설 수립과 검정하기

- 탐색적 데이터 분석은 **가설 (질문) 수립과 검정의 반복**으로 구성됨



# I 가설 수립과 검정하기 (예시)



- 문제 상황: 한 보험사에서 고객의 **이탈이 크게 늘고 있어**, 이탈할 고객들을 미리 **식별**하는 모델을 만들고, **이탈할 것이라 예상**되는 고객을 관리하여 이탈율을 줄이고자 함

- 가장 먼저 선행되어야 하는 작업은 **고객의 이탈 원인을 파악**하는 것이고, 파악된 원인을 모델의 **특징**으로 사용해야 함

| 가설 예시 | 검정 방법 |
|---|---|
| 보험 가입 기간이 긴 고객일수록 이탈율이 줄어들 것이다. | 보험 가입 기간과 고객 이탈 여부 간 t 검정 및 박스 플롯 시각화 |
| 고객 여정과 이탈율은 관계가 있을 것이다. | 주요 고객 여정 추출 및 주요 고객 여정 여부와 이탈율 간 카이 제곱 검정 및 히트맵 시각화 |

# I 주요 활동

| 구분 | 활동 내용 |
| --- | --- |
| 분석 목적 및 변수 확인 | • 대략적인 문제 정의<br>• 변수별 의미를 코드북 및 도메인 지식을 통해 확인 |
| 데이터 전체적으로 살펴보기 | • 데이터 자체에 문제가 없는지를 확인<br>• 데이터의 일부를 샘플링하여 하나하나 살펴보기<br>• 이상치 및 결측치 탐색<br>• 군집화 및 빈발 패턴 추출등을 통한 주요 패턴 파악 |
| 개별 속성값 확인 | • 개별 변수에 대한 빈도 분석 및 기술 통계<br>• 그래프 시각화 (히스토그램, 파이 차트, 박스 플롯)<br>• 분포 적합성 검정 |
| 속성 간 관계 파악 | • 카이 제곱 검정<br>• 상관 관계 분석<br>• t-검정 및 일원분산분석<br>• 산점도 및 히트맵 시각화 |

Chapter. 08

## 시작에 앞서: 탐색적 데이터 분석이란?

# 사례 소개

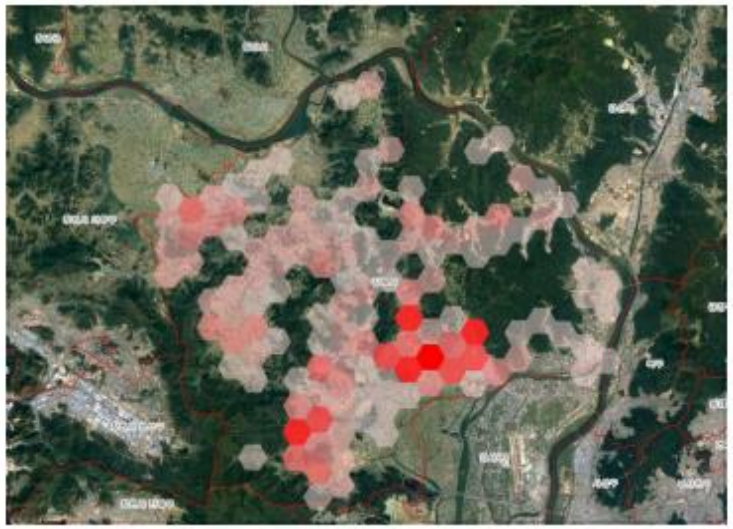
FAST CAMPUS
ONLINE
데이터 탐색과 전처리 I

강사. 안길승

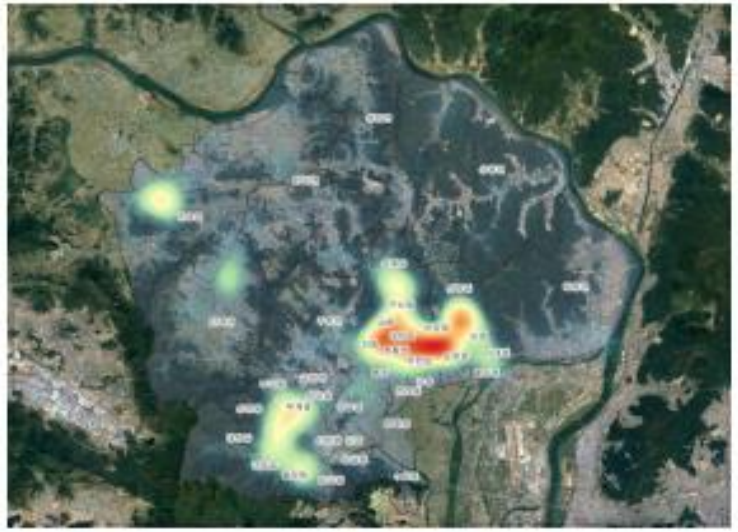
# I 사례 1. 김해시 화재 예측

- 건물별 화재 예측을 하는데 필요한 특징을 선정하는데 탐색적 데이터 분석을 수행함

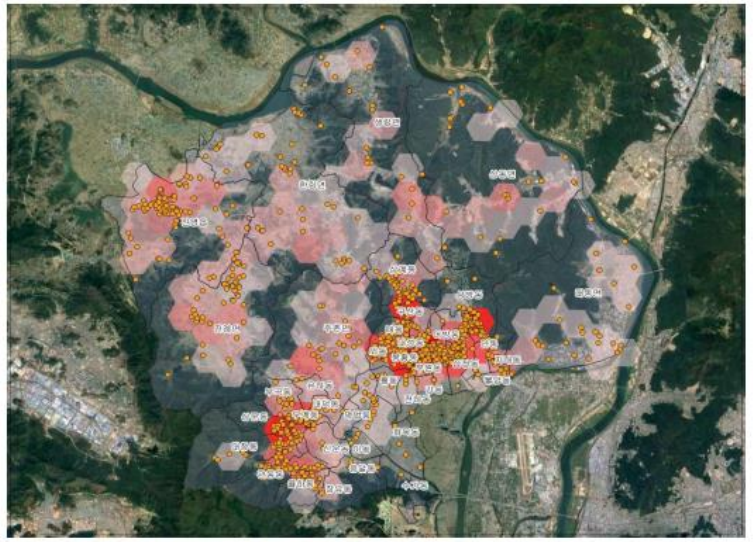
- 다소 특이한 특징으로 건축물 근처의 CCTV 현황을 사용하였는데, 그 근거는 화재 발생 빈도와 CCTV 현황 히트맵이 중첩된다는 것이었음



화재 발생 빈도 히트맵



CCTV 현황 히트맵

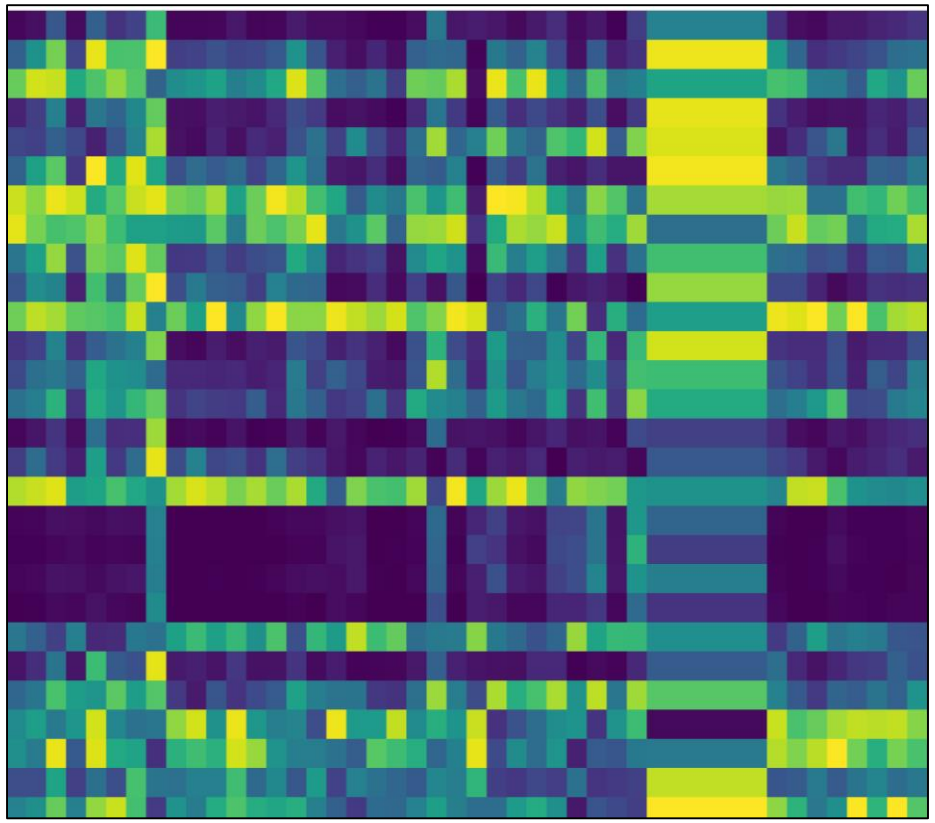


화재 발생 빈도 히트맵과 CCTV 현황 히트맵의 중첩 결과

출처: “빅데이터 기반의 김해시 화재발생 예측모델 개발”, 스마트시티 빅데이터 팀, LH

# I 사례 2. 밸브의 불량이 발생하는 공정상의 원인 파악

- 밸브의 생산 공정에서 발생하는 불량이 발생하는 원인을 파악하기 위해, 탐색적 데이터 분석을 수행함



제품 스펙과 공정 환경 간 상관 분석 결과

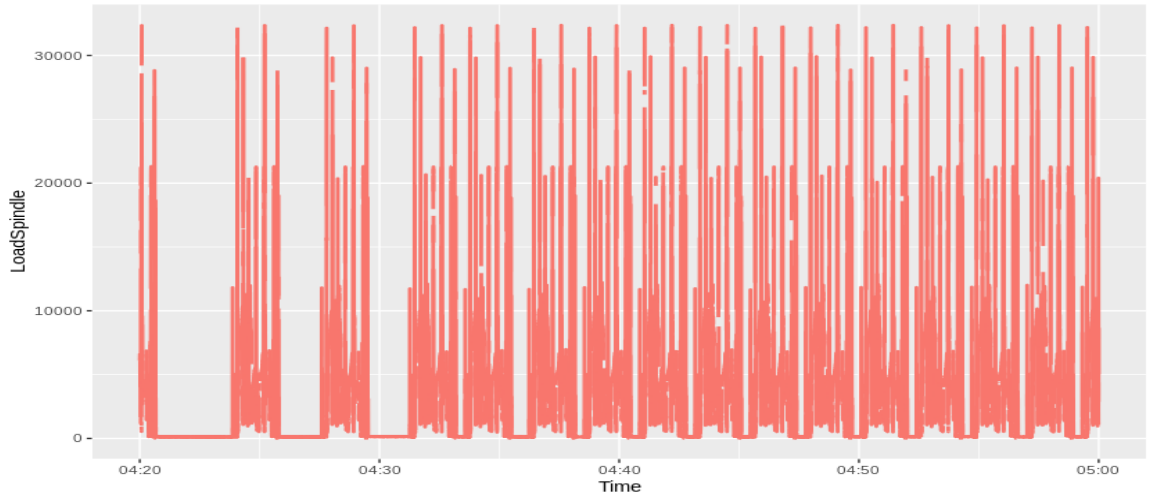
```
금형온도- ~ 사출즉시-바닥내경(33.27±0.13)-25A바닥내경

if 금형온도-하판(슬리브)-캐비티2 <= 40.6:
    사출즉시-바닥내경(33.27±0.13)-25A바닥내경 = 33.3625
else:
    if 금형온도-상판(코아)-캐비티4 <= 44.55:
        if 금형온도-상판(코아)-캐비티3 <= 44.35:
            if 금형온도-하판(슬리브)-캐비티1 <= 39.25:
                사출즉시-바닥내경(33.27±0.13)-25A바닥내경 = 33.34
            else:
                사출즉시-바닥내경(33.27±0.13)-25A바닥내경 = 33.3525
        else:
            if 금형온도-하판(코아)-캐비티1 <= 45.0:
                사출즉시-바닥내경(33.27±0.13)-25A바닥내경 = 33.3588
            else:
                사출즉시-바닥내경(33.27±0.13)-25A바닥내경 = 33.3825
    else:
        if 금형온도-상판(코아)-캐비티1 <= 46.0:
            if 금형온도-상판(코아)-캐비티3 <= 45.4:
                사출즉시-바닥내경(33.27±0.13)-25A바닥내경 = 33.3544
            else:
                사출즉시-바닥내경(33.27±0.13)-25A바닥내경 = 33.3438
        else:
            if 금형온도-상판(코아)-캐비티1 <= 47.7:
                사출즉시-바닥내경(33.27±0.13)-25A바닥내경 = 33.3325
            else:
                사출즉시-바닥내경(33.27±0.13)-25A바닥내경 = 33.355
```

의사결정나무 분석 결과

# I 사례 3. 설비 문제 알람의 원인 탐색



- 설비의 문제 알람이 발생하는 원인을 탐색하고, 설비 데이터 자체에 문제를 파악하기 위해 EDA를 적용함
- 분석 내용: 설비 데이터에 **비정상적으로 수집**되는 사례가 다수 있음을 확인함

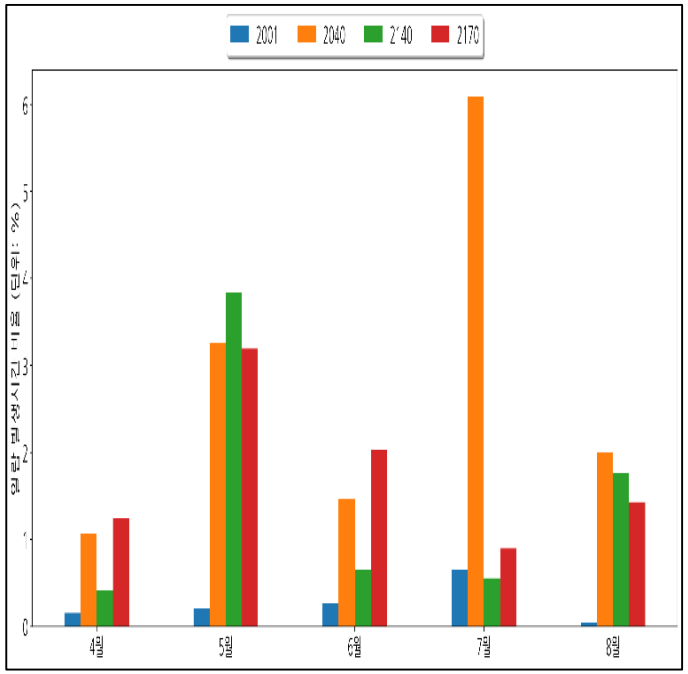


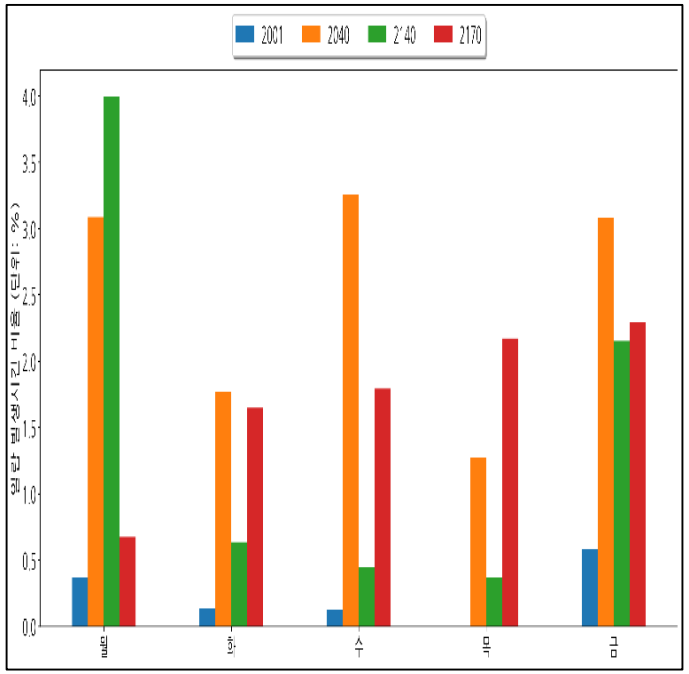

알람이 미발생했음에도 불구하고, 센서 값이
비정상적으로 들어오고 툴 코드는 정상적으로 들어옴

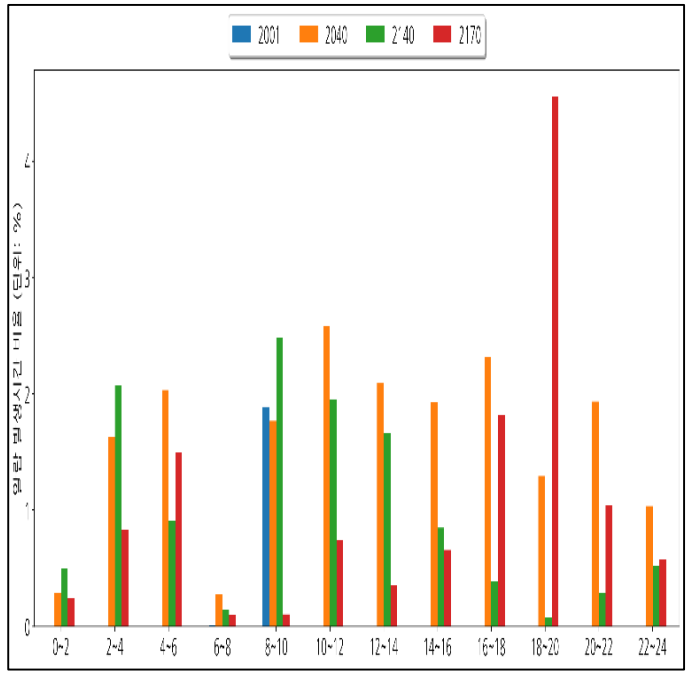
- 결론: 비정상적으로 수집된 데이터가 다수 있어, **재수집**이 필요함

# I 사례 3. 설비 문제 알람의 원인 탐색 (계속)

- 분석 내용 및 결과: 설비와 시간대별 알람 비율을 확인했으며, 비정상적인 구간이 존재함을 확인



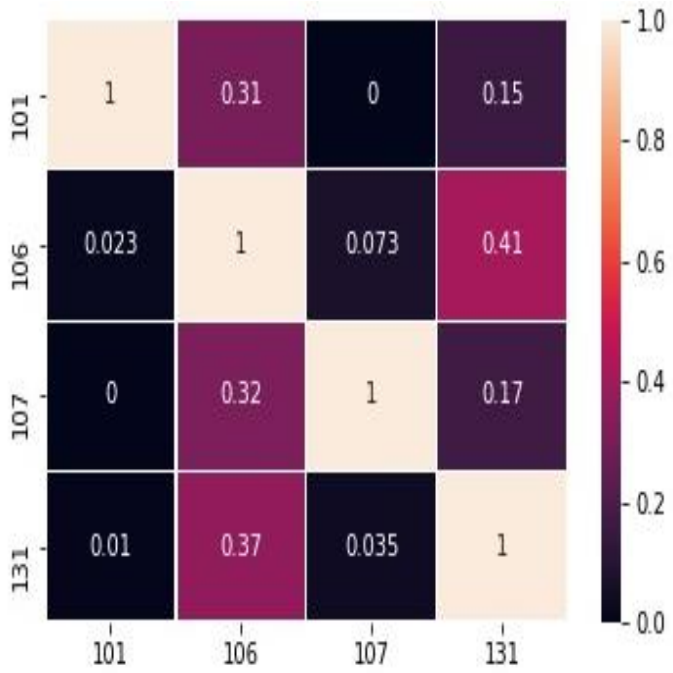




- 결론: 월, 요일, 시간대별 특정 이벤트가 존재한다는 가설을 세워 추가 탐색을 수행

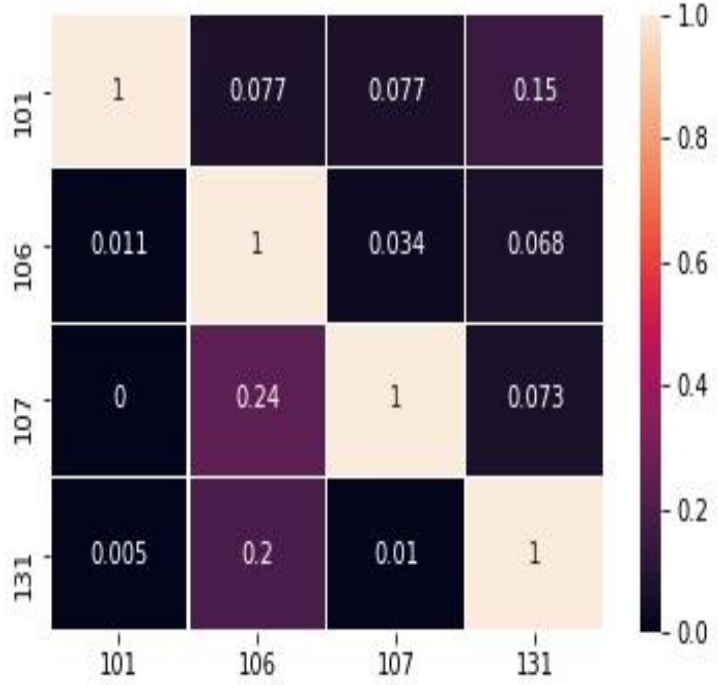
# I 사례 3. 설비 문제 알람의 원인 탐색 (계속)

- 분석 내용: 알람 간 동시 발생 여부 및 알람 간 선행 여부 분석

알람 간 동시 발생 여부




알람 간 선행 여부




- 결론: 동시 발생하는 알람과 선행 및 후행하는 알람 간 관계를 파악하여, 알람들을 카테고리화함