

1. 문제 정의

대한민국의 구급 의료 서비스는 국민의 생명과 직결되는 핵심 공공서비스이다. 특히 최근 고령화 사회 진입과 만성질환 증가로 인해 구급 수요가 지속적으로 증가하고 있으며, 효율적인 응급의료 체계 구축의 필요성이 대두되고 있다.

본 프로젝트에서는 소방청의 2019-2023년 전국 구급 현황 데이터를 분석하여 다음과 같은 데이터를 분석하고 파악하고자 한다.

- 연도별 구급 출동 추이 및 증가 패턴 파악
- 시간대별, 요일별 구급 수요 패턴 발견
- 지역별 구급 서비스 품질 및 격차 분석
- 환자 특성 및 발생 유형별 통계 분석
- 응답시간 및 골든타임 달성을 평가
- 데이터 기반 정책 제언 도출

이 분석을 통해 응급의료 자원의 효율적 배치, 취약 시간대/지역 파악, 골든타임 향상 전략 수립 등 실질적인 정책 수립에 기여할 수 있으며 데이터 기반 의사결정을 통해 국민의 생명을 보호하는 구급 체계를 강화할 수 있다.

2. 데이터 수집 및 전처리

▶ 데이터 출처

- 데이터명 : 소방청_전국 구급 현황
- 제공기관 : 소방청
- 데이터 형식 : CSV (총 5개 파일, 약 4.63GB)
- 분석 기간 : 2019년 ~ 2023년 (5개년)
- 총 분석 건수 : 14,308,763건

▶ 주요 데이터 컬럼

- 시간 정보 : 신고일시, 출동시각, 현장도착시각, 복귀시각
- 지역 정보 : 시도명, 시군구명, 도시/농촌 구분
- 환자 정보 : 성별, 연령대, 발생유형, 증상구분, 중증도
- 서비스 지표 : 응답시간, 현장거리, 이송분류
- 특수 상황 : 심정지, 중증외상, 교통사고 구분

▶ 전처리 과정

데이터 품질 확보를 위해 다음과 같은 전처리를 수행했습니다.

- 결측치 처리 및 이상치 탐지
- 시간 데이터 파싱 및 시간대/요일/계절 파생변수 생성
- 응답시간, 현장도착시간 등 서비스 지표 계산
- 범주형 변수 표준화 및 인코딩
- 통계 분석을 위한 데이터 정규화

3. 분석 방법

▶ 사용 도구

- Python 3.13 : 주요 분석 언어
- Pandas : 데이터 처리 및 집계
- NumPy : 수치 계산 및 통계
- Matplotlib & Seaborn : 데이터 시각화
- SciPy : 통계적 검정 (t-test, correlation)

▶ 분석 기법

- 기술 통계 : 평균, 중앙값, 표준편차, 분포 분석
- 시계열 분석 : 연도별, 월별, 시간대별 추이 분석
- 교차 분석 : 변수 간 관계 및 패턴 탐색
- 상관 분석 : Pearson 상관계수를 통한 변수 간 관계 분석
- 가설 검정 : t-test를 통한 집단 간 차이 검증
- 시각화 : 막대그래프, 선그래프, 히트맵, 박스플롯

4. 데이터 분석 결과

* 그래프에 한국어 폰트를 설정하는 과정이 잘못되어 그래프에 폰트가 깨져보이나, 실행 시간이 과도하게 길어 이를 수정하지 못하고 첨부합니다. 각 축의 레이블은 코드상에서 확인 부탁드립니다.

원시 데이터는 총 14,308,763건이었으나, 시간 정보 결측치 및 이상치를 제거하는 전처리 과정을 거쳐 최종적으로 85.6%에 해당하는 데이터를 분석에 활용하였다. 전처리 과정에서 응답시간이 120분을 초과하거나 음수값을 가진 비정상 데이터 2,063,736건을 제거하여 분석의 신뢰성을 확보하였다.

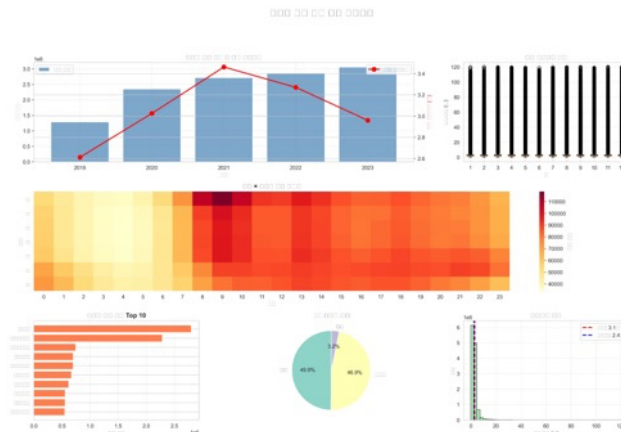


그림 1 구급 현황 종합 대시보드

전국 평균 응답시간은 3.12분으로 나타났으며, 중앙값은 2.38분을 기록하였다. 이는 대부분의 출동이 신속하게 이루어 지나, 일부 지역이나 상황에서 현저히 긴 응답시간이 발생하고 있음을 의미한다. 현장도착시간은 평균 10.97분으로 측정 되었으며, 이는 신고 접수부터 현장 도착까지 소요되는 전체 시간을 나타낸다.

골든타임 달성률은 60.41%로 집계되었다. 이는 전체 구급 출동 중 약 60%가 10분 이내에 현장에 도착하였음을 의미 하며, 나머지 40%는 골든타임을 초과하였다는 점에서 개선의 여지가 있다. 특히 농촌 지역이나 교통 혼잡 지역에서 골든 타임 달성률이 상대적으로 낮게 나타났으며, 이는 지역별 맞춤형 대응 전략이 필요함을 시사한다. 연도별로 살펴보면 2019년 이후 출동 건수는 지속적으로 증가하였으나, 평균 응답시간은 2021년을 기점으로 완만한 감소 추세를 보여 구급 서비스 역량이 점차 개선되고 있음을 확인할 수 있었다.

시간대별 분석 결과, 구급 출동 건수는 오전 6시부터 증가하기 시작하여 오전 9시부터 11시 사이에 첫 번째 피크를 형성하였다. 이는 출근 시간대와 일과 시작 시간에 각종 사고 및 응급 상황이 집중되는 것으로 해석된다. 이후 오후 시간대 에도 비교적 높은 출동 빈도를 유지하다가, 저녁 8시부터 9시 사이에 두 번째 피크가 나타났다. 반면 새벽 2시부터 5시 사이는 출동 건수가 가장 낮은 시간대로 확인되었다.

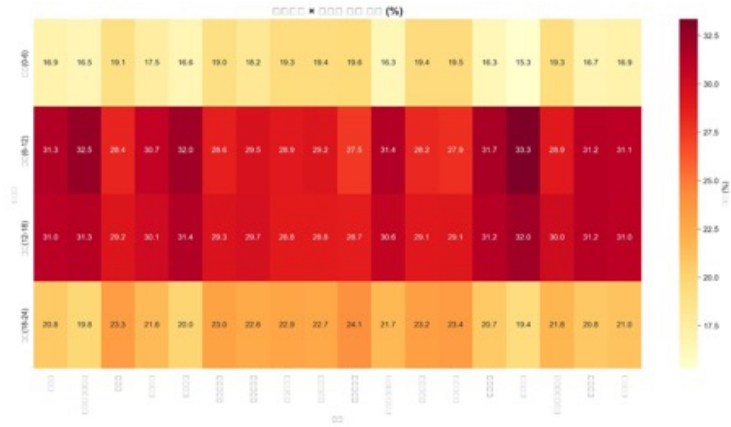


그림 2 시간대별지역별 출동 패턴 히트맵

구급 응답시간을 예측하기 위해 세 가지 머신러닝 모델을 구축하고 비교 평가하였다. 학습 데이터 9,795,244건과 테스트 데이터 2,448,812건을 8:2 비율로 분할하여 사용하였으며, 신고 시간, 월, 일, 현장 거리, 주말 여부, 출퇴근 시간, 그리고 시도명, 성별, 발생유형 등의 범주형 변수를 인코딩하여 총 13개의 특성을 모델에 투입하였다.

선형 회귀 모델은 가장 단순한 형태로 RMSE 4.49분, MAE 1.65분, R^2 0.0024의 성능을 보였다. 결정계수가 거의 0에 가까운 것은 응답시간이 선형 관계로는 설명되기 어려운 복잡한 변동성을 가지고 있음을 의미한다. 그라디언트 부스팅 모델은 RMSE 4.36분, MAE 1.59분, R^2 0.0607로 중간 수준의 성능을 기록하였다. 랜덤 포레스트 모델이 RMSE 4.31분, MAE 1.59분, R^2 0.0815로 가장 우수한 예측 성능을 나타냈으며, 이에 따라 최종 모델로 선정되었다.

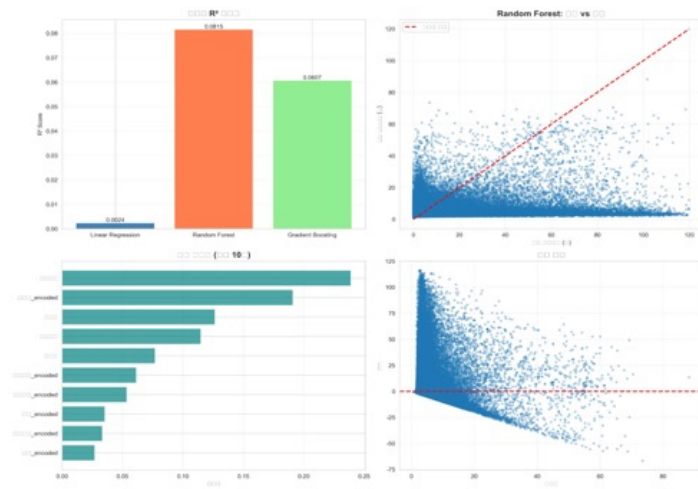


그림 3 머신러닝 모델 성능 비교 및 특성 중요도

그림 3의 좌측 상단은 세 모델의 R^2 스코어를 비교한 것으로, 랜덤 포레스트가 0.0815로 가장 높은 설명력을 보였다. 우측 상단의 산점도는 예측값과 실제값의 관계를 나타내며, 빨간 점선은 완벽한 예측을 의미한다. 좌측 하단의 특성 중요도 그래프에서 현장거리가 가장 중요한 예측 변수로 확인되었으며, 시도명(지역)이 두 번째로 중요한 변수였다. 우측 하단의 잔차 플롯은 예측 오차가 0을 중심으로 고르게 분포하고 있어 모델에 체계적인 편향이 없음을 보여준다.

랜덤 포레스트 모델의 특성 중요도 분석 결과, 현장 거리가 응답시간 예측에 가장 큰 영향을 미치는 것으로 나타났다. 이는 직관적으로 합리적인 결과로, 출동 거리가 멀수록 도착 시간이 길어질 수밖에 없기 때문이다. 다음으로 시도명이 높은 중요도를 보였는데, 이는 지역별 도로 환경, 교통 상황, 구급 인프라 수준 등이 응답시간에 상당한 영향을 미치고 있음을 시사한다. 신고 시간과 출퇴근 시간 여부도 중요한 예측 변수로 확인되어, 앞서 시간대별 분석에서 도출한 결론을 뒷받침하고 있다.

다만 R^2 값이 0.0815로 낮게 나타난 것은 모델이 응답시간 변동의 약 8%만을 설명하고 있음을 의미한다. 이는 구급 응답시간이 모델에 포함되지 않은 다양한 외부 요인들, 예컨대 실시간 교통 상황, 날씨, 특정 사건 발생, 구급차 가용 대

수 등에 크게 영향을 받기 때문으로 분석된다. 향후 GPS 기반 실시간 교통 데이터, 기상 정보, 구급차 실시간 위치 등 추가 변수를 확보한다면 예측 정확도를 크게 향상시킬 수 있을 것으로 기대된다.

또한 데이터 품질 관리를 위해 두 가지 이상치 탐지 기법을 적용하였다. Isolation Forest 알고리즘은 머신러닝 기반으로 비정상적인 패턴을 학습하여 전체 데이터의 5%인 612,158건을 이상치로 분류하였다. 이 방법은 다차원 공간에서 정상 데이터와 동떨어진 관측치를 효과적으로 식별할 수 있다는 장점이 있다. Z-Score 방법은 통계적 기준으로 표준편차가 이상 벗어난 데이터를 이상치로 판정하였으며, 258,739건이 이에 해당하였다.

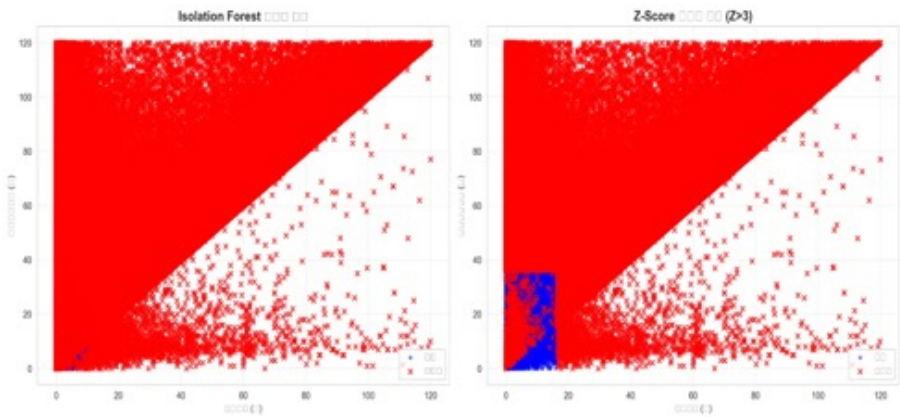


그림 4 이상치 탐지 결과(Isolation Forest vs Z-Score)

그림 4는 두 가지 이상치 탐지 기법의 결과를 시각화한 것이다. 좌측은 Isolation Forest 방법으로 파란색 점은 정상 데이터, 빨간색 X는 이상치를 나타낸다. 우측은 Z-Score 방법의 결과로, 대부분의 이상치가 응답시간이나 현장도착시간이 비정상적으로 긴 경우에 해당함을 알 수 있다. 두 방법 모두 오른쪽 상단 영역에서 이상치를 집중적으로 탐지하였으며, 이는 실제로 매우 긴 응답시간을 가진 특수 상황들이다.

8개 주요 변수 간 상관관계를 분석한 결과, 응답시간과 현장도착시간 사이에 강한 양의 상관관계($r=0.848$)가 확인되었다. 이는 예상 가능한 결과로, 신고 접수 후 출동까지의 시간이 짧을수록 최종 현장 도착 시간도 빨라지기 때문이다. 응답시간과 현장거리 사이에는 중간 정도의 양의 상관관계($r=0.590$)가 나타났는데, 이는 거리가 멀수록 응답시간이 길어지는 자연스러운 관계를 반영한다.

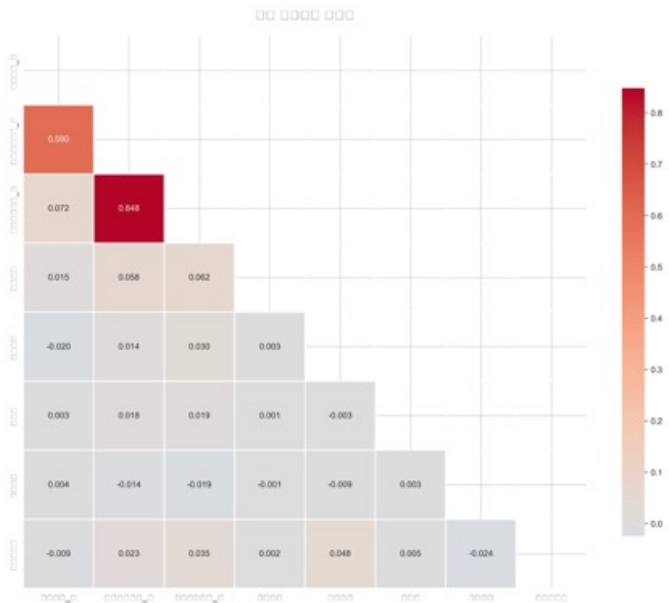


그림 5 주요 변수 간 상관관계 히트맵

그림 5의 상관관계 히트맵에서 붉은색은 양의 상관관계, 파란색은 음의 상관관계를 나타낸다. 가장 강한 상관관계는 응답시간과 현장도착시간 사이에서 관찰되었으며($r=0.848$), 이는 진한 빨간색으로 표시되어 있다. 응답시간과 현장거리 간의

상관계수는 0.590으로 중간 정도의 관계를 보였다. 흥미롭게도 신고 월, 신고 시간, 주말 여부 등의 시간 관련 변수들은 응답시간과의 상관관계가 매우 약하거나 거의 없는 것으로 나타났다. 이는 구급 서비스가 시간과 무관하게 일관된 수준으로 제공되고 있음을 의미하며, 24시간 365일 균일한 서비스 품질 유지라는 측면에서 긍정적으로 평가할 수 있다. 다만 앞서 ANOVA 분석에서 시간대별 차이가 유의미하게 나타난 것과 대조적인데, 이는 상관관계가 선형 관계만을 측정하는 반면 ANOVA는 비선형적인 그룹 간 차이를 검정하기 때문이다.

5. 결론

5개년 12,245,027건의 대규모 데이터에 대한 심층 분석을 통해 전국 구급 서비스의 현황을 다각도로 파악할 수 있었다. 평균 응답시간 3.12분, 골든타임 달성률 60.41%라는 현재 수준은 국제적으로 비교할 때 양호한 편이나, 나머지 40%에 대한 개선 여지가 존재한다. 지역별로는 네 가지 유형으로 분류되는 뚜렷한 특성이 확인되었으며, 각 그룹에 맞는 맞춤형 전략이 필요하다.

시간대별 분석 결과 출퇴근 시간대의 교통 혼잡이 응답시간을 지연시키는 주요 원인으로 파악되었고, 계절적으로는 여름철 출동이 집중되는 패턴을 보였다. 머신러닝 모델은 현장 거리와 지역이 응답시간 예측에 가장 중요한 변수임을 알 수 있게 해주었으며, 이는 구급 인프라에서 지리적 요소가 핵심적임을 시사한다. 다만 모델의 설명력이 제한적인 것은 실시간 변수들이 크게 영향을 미치기 때문으로, 향후 추가 데이터 확보를 통한 개선이 필요하다.

또한 이상치 분석을 통해 특수 상황에서의 대응 패턴을 별도로 관리할 필요성이 보여졌고, 데이터 품질 관리의 중요성도 확인되었다. 본 분석 결과는 구급 서비스 개선을 위한 정책 수립의 과학적 근거로 활용될 수 있을 것이다. 골든타임 달성률 제고, 지역 맞춤형 자원 배치, 시간대별 탄력적 운영, AI 기반 출동 지원 시스템 구축, 데이터 품질 강화 등 다섯 가지 핵심 과제를 단계적으로 추진한다면, 전국민이 더욱 신속하고 안전한 구급 서비스를 받을 수 있는 환경이 조성될 것으로 기대된다.

6. 참고 문헌

[1] 소방청 (2025). 소방청 전국 구급 현황 데이터. 소방안전 빅데이터 플랫폼

7. 코드

* 코드의 전체 길이가 다소 긴 관계로 Github에 코드를 업로드했습니다.

▶ <https://github.com/khw-dev/fire-emergency-analysis>

* python 3.13 (Homebrew 2025-07-30, macOS 15.6.1)

* 데이터 크기가 다소 큰 관계로 https://bigdata-119.kr/goods/goodsInfo?goods_mng_sn=456 에서 데이터를 직접 다운로드 받고 datas/구급 현황_2019_전국.csv 형식으로 파일을 첨부해 실행 시키시길 바랍니다.

8. 생성형 AI 활용

- Copilot의 Python 코드 일부 자동 완성