문제발굴 신청서

2026 DATA·AI 분석 경진대회

1. 문제 개요

○ 문제 제목

프로야구 홈경기 정보를 활용한 지역 상권 일별 수요 예측

○ 배경 및 목적

지방자치단체는 프로구단 연고 유치와 경기장 건설에 큰 규모의 공공재정을 투입하면서 '지역경제 활성화'를 주요 근거로 제시한다. 대전광역시도 2025년 3월 5일 2만 석 규모의 대전 한화생명 볼파크를 개장했다.

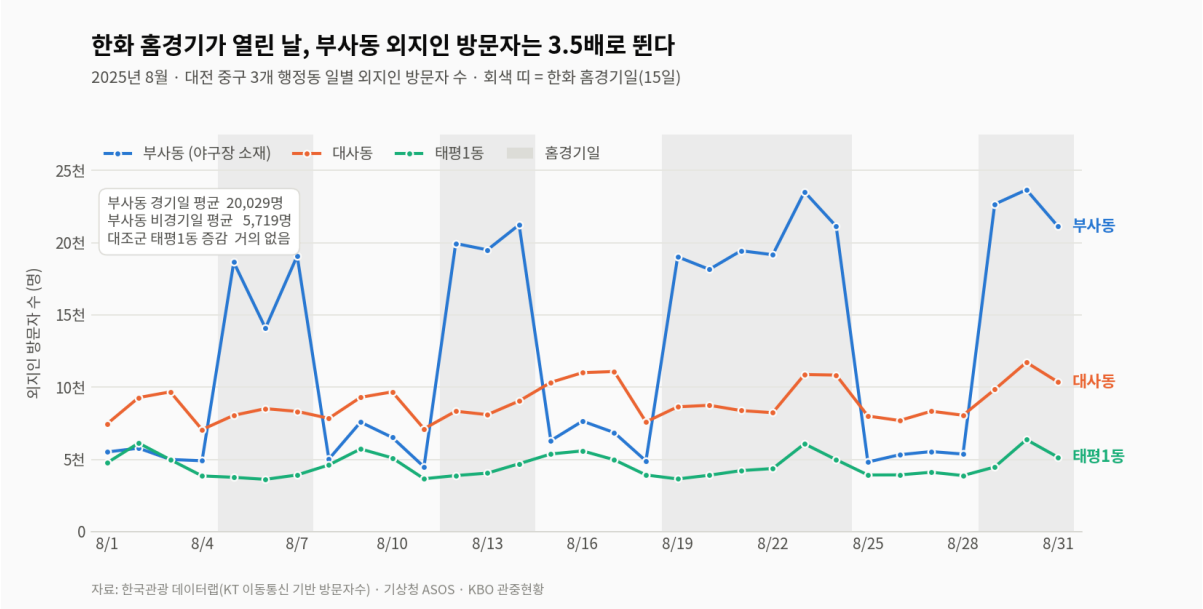
그런데 홈경기가 주변 상권에 실제로 얼마만큼의 수요를 만드는지는 잘 알려져 있지 않다. 기존 분석은 대부분 건설 이전에 추정한 값이거나 개장 이후의 총액 수치다. 그래서 셔틀 증차나 야간 영업 지원 같은 경기일 연계 정책도, 소상공인의 재고·인력 운영도 대체로 경험에 의존해 이루어진다.

이 문제는 경기 일정처럼 미리 알 수 있는 정보만으로 특정 지역의 일별 상권 수요를 예측하는 모델을 만드는 것을 목표로 한다.

○ 문제 설명

제출한 샘플 데이터(2025년 8월, 대전 중구 3개 행정동)에서 이미 신호가 뚜렷하게 보인다. 야구장이 있는 부사동의 외지인 방문자는 홈경기일에 평균 20,029명, 경기가 없는 날에 5,719명으로 3.5배 차이가 났다.

이 차이가 계절이나 요일 때문이 아니라는 것도 같이 확인할 수 있었다. 같은 기간 대조군인 태평1동은 경기일에 오히려 183명 줄었고, 바로 옆 대사동도 480명 느는 데 그쳤다. 이중차분으로 계산하면 순효과가 14,493명이고, 요일과 기온·강수량을 넣고 회귀분석을 해도 홈경기 하루당 약 14,800명이 나온다. 관중 수와 방문자 수의 상관계수는 0.804였다.



[그림] 2025년 8월 대전 중구 3개 행정동 일별 외지인 방문자 수 (회색 띠 = 한화 홈경기일)

증가분의 대부분은 외지인이었다. 현지인 방문자는 1.2배 늘어나는 데 그쳤으니, 늘어난 수요는 동네 사람이 더 나온 것이 아니라 밖에서 들어온 것으로 볼 수 있다.

다만 이 수치들은 31일치 표본에서 나온 것이라 그대로 일반화하기는 어렵다. 관측치에 비해 변수가 많아 과적합 가능성이 있고, 관중 수도 15경기 중 9경기가 매진 상한인 17,000명으로 같아서 관중 규모에 따른 차이를 충분히 보기 어렵다. 참가자는 시즌 전체로 기간을 넓혀 이 관계가 유지되는지 검증하고, 예측 모델로 일반화해야 한다.

기간을 넓히면 확인할 수 있는 것이 더 있다. 홈경기는 한 시즌에 70여 회 반복되므로 관측치가 충분히 쌓이고, 2025년 3월 신축 구장 개장이라는 변화 시점이 있어 개장 전후 비교도 가능해진다.

2. 목표

⬭ 문제 해결 목표 성능 또는 기대 결과물에 대한 설명

- 테스트 기간에 대한 일자별 상권 수요(외지인 방문자 수) 예측값

- 경기 정보를 사용하지 않은 베이스라인 대비 예측오차 개선

- 어떤 조건에서 효과가 커지는지에 대한 해석 (관중 규모, 요일, 경기 시작 시각 등)

- 효과의 공간적 범위(경기장 소재 동 대비 인접 동 파급 정도) 추정

⬭ 평가 방법 및 설명

- 주 평가지표 : RMSE (보조지표 MAPE)

- 베이스라인 : 요일 정보만 사용하는 모델. 샘플 데이터 기준 RMSE 6,524명(평균 대비 51.6%)

- 참고 수준 : 홈경기 여부만 추가해도 RMSE가 1,159명(평균 대비 9.2%)까지 떨어진다. 다만 31일 표본 기준이므로 참가자는 더 긴 기간에서 이 수준이 유지되는지 확인하고, 관중 규모와 시간대를 활용해 정확도와 설명력을 더 확보해야 한다.

- 해석 가능성 : 변수 중요도 또는 회귀계수에 대한 설명 제시 여부

3. 데이터 설명

⬭ 데이터명

프로야구 홈경기-지역 상권 일별 결합 데이터셋 (대전 중구 3개 행정동, 2025년 8월)

⬭ 유형**,** 포맷**,** 용량**/**건수 등

정형 데이터 / CSV(UTF-8) / 93행 × 16열 (3개 행정동 × 31일), 결측 0건

⬭ 문제해결을 위해 필요한 설명 **(**데이터 사전**)**

속성	설명	데이터 타입	값 범위 혹은 범주	참고
date	일자	String	2025-08-01 ~ 2025-08-31	
sigungu_code	시군구 코드	String	30110	대전광역시 중구
sigungu_name	시군구명	String	대전광역시 중구	
dong	행정동명	String	부사동, 대사동, 태평1동	
is_stadium_dong	야구장 소재 여부	Integer	0, 1	부사동=1 (처치군)

dow	요일	String	월, 화, 수, 목, 금, 토, 일	
is_holiday	공휴일 여부	Integer	0, 1	광복절
is_gameday	홈경기 개최 여부	Integer	0, 1	핵심 설명변수
attendance	경기 관중 수	Integer	0, 13919 ~ 17000	비경기일은 0
opponent	상대팀	String	KT, 롯데, 두산, SSG, 삼성	비경기일은 공란
start_time	경기 시작 시각	String	18:00, 18:30	비경기일은 공란
temp_avg	일평균 기온	Float	24.3 ~ 31.2 (℃)	기상청 대전(133)
precip	일강수량	Float	0 ~ 46.7 (mm)	기상청 대전(133)
visitor_out	외지인 방문자 수	Float	3,616 ~ 23,667	예측 대상
visitor_local	현지인 방문자 수	Float	4,956 ~ 9,331	
visitor_total	전체 방문자 수	Float	9,950 ~ 32,000	visitor_out + visitor_local

ㅇ 기타 참고 데이터에 대한 설명

- 한국관광 데이터랩(한국관광공사) : 행정동 단위 일별 방문자 수. 내국인은 KT 이동통신 데이터 기반

- 기상청 기상자료개방포털 : 종관기상관측(ASOS) 대전 지점(133) 일자료

- KBO 공식 기록실 : 경기 일정, 경기별 관중 수, 상대팀, 경기 시작 시각

- 유의사항 : 방문자 수는 이동통신 표본을 모수 추정한 값이므로 절대값보다 추이와 비율로 해석해야 한다. 이 문제는 경기일과 비경기일의 상대적 차이를 다루므로 그 성격에 부합한다.

4. 데이터 샘플

2025년 8월 한 달, 대전 중구 3개 행정동(부사동·대사동·태평1동)의 일별 데이터 93행을 제출한다. 처치군(부사동)과 대조군(태평1동), 인접 동(대사동)을 함께 포함해 이중차분 설계가 곧바로 가능하도록 구성했다.

문제 채택 이후에는 2025년 시즌 전체(3~10월, 약 700행)와 타 구단 연고지로 확대하여 제공한다.

첨부 파일 : sample_data_final.csv

5. 제출 형식

ㅇ 참가자로부터 받을 결과 파일의 형식

예측 결과 파일 (CSV)

date	dong	pred_visitor_out
2025-08-01	부사동	5480
2025-08-05	부사동	19870

분석 보고서 (PDF) : 모델 구조, 변수 선택 근거, 경기 효과에 대한 해석 포함