2조 프로젝트제안서

1.주제	데이터 분석을 통한 2024 야구 승률 예측
2. 주제 선정 배경	한국 프로야구는 수많은 팬들에게 큰 사랑을 받으며, 매 시즌 팀별 경기 결과와 성적 변화는 큰 관심의 대상이 되고 있다. 이에 따라 팀의 성적에 영향을 미치는 요소를 파악하고 나아가 향후 성적을 예측하는 연구가 활발해지고 있다. 이번 프로젝트를 통해 팀의 성적에 영향을 주는 요인을 직접 분석해 보는 경험을 하고자 하였고 요인들 중 주원인이라고 예상되는 '원정 간 이동거리'를 중심으로 '승률'과의 상관관계 분석을 시작하게 되었다. 이러한 분석은 전략적인 의사 결정 지원으로, 분석 결과가 충분한 의미를 가지게 되면 앞으로 각 팀의 이동 일정이나 경기 전후의 훈련 계획 등을 세밀하게 조정할 수 있게 된다. 이로써 단기적인 승률뿐 아니라 팀의 전반적인 성적 관리와 운영에도 도움이 될 것이다.
3. 활용 데이터	-KBO 사이트, 구단별 공식 기록 자료 웹크롤링 (2015~2023년 팀의 연도별 스케줄 데이터 수집) (팀별 타자, 투수, 수비 기록 데이터 수집) (연도별 승률 데이터 수집) (연도별 관중수 데이터 수집) -뉴스와 기사를 통한 데이터 수집 (연도별 팀 예산 데이터 수집)
4. 분석 내용	'이동하는 누적거리가 클 수록 승률이 낮다 '라는 세부적인 가설 하나를 두고 이동거리와 승률의 상관관계를 분석한 후 결과를 가설과 비교해보며 2차 목표로 요인을 확장해가면서 단계별로 진행하였다. <1차 목표 : 누적 이동거리와 승률의 변화> 1. 데이터 수집 1)팀의 연도별 스케줄 데이터 정확한 승률 차이를 확인하기 위해 2015년에서 2023년까지 10년치의 연도별 스케줄을 크롤링하여 팀의 이동 거리와 승패의 변수를 도출할 수 있는 기초 데이터를 확보하였다. 2)연도별 승률 데이터 KBO 홈페이지에서 10년치의 구단별 승률을 크롤링하였다. 2. 데이터 정제 1) 구단 및 홈구장 이름 통일 KBO 홈페이지에서 얻은 데이터에서 여러 구단이나 홈구장의 이름이 서로 다르게 표기 된 경우가 많아 이를 일관성 있게 통일하는 과정이 필요했다. 이 단계를 통해 이후 분석 단계에서 혼동 없이 데이터를 다룰 수 있게 하였다. 2) 이동거리 계산 및 누적거리 추가

구단별 홈구장과 원구장과의 거리를 구하여 이동거리를 계산하고 시즌이 진행되면서 누적되는 거리를 추가해 승률과의 비교를 심도 있게 분석할 수 있게 하였다.

3) 승패 결과 및 누적 승률 추가
본격적인 승률 분석에 핵심적인 자료로, 각 경기의 승패 결과와 누적 승률을 추출하여 추가한 후 이동 거리와 승률 간의 상관관계를 파악하는데 필요한 지표로 만들었다.

3. 시각화
1) folium 지도화
2) 파이썬

4. 분석 결과 및 문제점의 원인 분석
1)분석결과
시각화를 통해 확인한 결과 이동거리와 승률 간에 유의미한 상관관계를 발견하지 못하였다. 기대했던 것과 달리, 장거리 이동이 많은 경우에도 승률에 큰 변화가 나타나지 않거나, 일정한 패턴이 보이지 않는 팀이 많아 이동거리가 팀의 성적에 미치는 영향이 예상보다 크지 않다는 결론을 내리게 되었다.
2)원인 분석
(1) KIA의 2021년과 2024년 승률비교
누적 이동거리가 두 해 모두 비슷한 수준이었으나 승률 차이가 매우 크게 나타났다.
(1) 2015년과 2022년의 전체 승률과 누적 이동거리 비교
두 해의 누적 이동거리가 상당히 컸음에도 불구하고 승률 차이는 미미하게 나타났다.

<2차 목표 : 팀 성적에 영향을 미칠 수 있는 다양한 요인 분석>
1.데이터 수집
이동거리 외에 다른 변수들이 승률에 어떻게 영향을 미치는지 분석하기 위해 각 팀의 타자, 투수, 수비 등과 같은 세부 기록 데이터를 추가적으로 활용하기로 하였다.
1)팀별 타자, 투수, 수비 기록 데이터
2)연도별 승률 데이터
3)뉴스와 기사를 통한 연도별 팀 예산 데이터
4)연도별 관중수 현황 데이터

2. 데이터 정제
1) 결측값 제거
일부 결측값이 존재하여 이를 평균값으로 대체하였다.
2) 차원 축소
승패의 요인이 되는 연도별 다양한 변수나 특성들 중에서 핵심적인 정보만 남기고 상대적으로 덜 중요한 정보는 제거하여 변수의 수를 줄이는 작업을 하였다. 이를 위해 **주성분 분석(PCA)**라는 기법을 사용하였다.
3) 타겟 변수 및 피처 설정
타겟 변수는 팀의 승률로 설정하였고 피처로는 데이터에서 추출된 숫자형 변수들을 활용하였다.

3.시각화
1) 파이썬

4. 예측 모델 구현 및 모델 결과 분석
1) 팀의 승률 예측을 위한 회귀 모델 사용
(1) 선형 회귀(Linear Regression)
(2) 릿지 회귀(Ridge Regression)와 라쏘 회귀(Lasso Regression)
(3) 서포트 벡터 회귀(SVR)
(4) 결정 트리 회귀(Decision Tree Regressor)
(5) KNN 회귀

	(6) 랜덤 포레스트 회귀(Random Forest Regressor) (7) 그래디언트 부스팅 회귀(Gradient Boosting Regressor) (8) XGBoost, LightGBM, CatBoost 회귀 모델 2) 모델 결과 분석 (1) 데이터 셋 구성 : 2015~2023까지 9년의 기록을 사용 (2024년 데이터는 미래 예측을 위해 예외시켰다.) (2) 모델 예측 결과 대부분의 팀에서 실제 승률과 유사한 예측 승률을 도출하였다. (3) 오차 범위 : 최소 0.001에서 최대 0.07
5.결론 및 시사점	첫 번째 목표였던 이동거리와 승률의 관계에서는 유의미한 상관관계를 발견하지 못했고, 팀의 이동거리가 승률에 직접적인 영향을 미치지 않는다는 결론에 도달했으며, 이 결과는 팀 성적에 영향을 미치는 더 중요한 요인들이 존재할 것이라는 가능성을 시사했다. 이후 목표를 수정하여 팀의 기록 데이터를 활용한 승률 예측에 집중해 성적 예측을 수행한 결과, 팀 성적에 영향을 미치는 여러 요인을 성공적으로 분석할 수 있었고 모델을 적용시켜 각 팀의 성적을 정확하게 예측할 수 있게 되었다. 프로젝트를 진행하면서 나온 한계점 및 개선방안으로는 데이터의 양적 한계, 오차 원인 분석, 추가 변수 고려를 생각해 볼 수 있다. 유의미한 장기적 패턴을 분석하기에 다소 데이터의 양이 적었고 더욱 안정적이고 신뢰적인 결과를 얻기 위해선 더 많은 연도의 데이터를 추가할 필요가 있을 것 같다. 또한 모델의 정확성을 개선하기 위한 오차의 원인을 분석하지 못한 것이 한계라고 볼 수 있겠다. 현재 분석에는 정량적 변수들이 중심이 되었으나 감독 교체, 선수 이적, 주요 선수의 부상과 같은 비정량적 요인도 성적에 중요한 영향을 줄 수 있는데 앞으로는 숫자로 나타내기 어려운 요소들을 지표로 만들어 데이터로 활용해 볼 필요도 있을 것 같다. 이번 프로젝트를 통해 웹크롤링을 활용한 데이터 수집과 파이썬으로 정제하는 기초 작업부터 결과를 한 눈에 분석하는 시각화, folium을 경험할 수 있었다. 더 나아가 고도화된 모델 적용과 예측 결과 분석까지 모두 다룰 수 있는 시간이었다. 또한 프로젝트 진행 중 목표를 유연하게 조정함으로써 문제 해결 능력과 목표 설정의 중요성을 깊이 인식하게 되었다. 이는 향후 데이터 분석 프로젝트를 수행하는 데 있어 매우 중요한 배움이 되었다.
6. 출처 및 기타	1.참고 문헌 1) 머신러닝 알고리즘 및 라이브러리 문서 2) 관련 연구 논문 및 기사 변수의 유형: https://m.blog.naver.com/gksshdk8003/222404187056 PCA란 무엇인가 :https://www.ibm.com/kr-ko/topics/principal-component-analysis 2. 데이터 출처 1) KBO 공식 홈페이지 2) 구단별 공식 기록 자료