슬라이드3

목적

이 프로젝트의 첫 번째 목적은 야구 팀이 원정 경기에서 이동하는 거리와 경기 승률 사이의 관계를 알아보는 것입니다. 즉, 팀이 장거리를 이동했을 때 승률에 어떤 변화가 있는지를 분석해 보고자 했습니다. 이를 통해 승률에 영향을 미치는 요인으로 이동거리가 과연 어느 정도로 중요한지 파악하는 것이 목표입니다.

두 번째로는 이동거리 외에도 팀의 성적에 영향을 줄 수 있는 다양한 요소들을 찾고자 했습니다. 예를 들어, 경기 스케줄이나 기후, 선수들의 피로도 등이 어떤 식으로 성적에 작용하는지 이해하려는 것이죠.

배경

왜 이동거리와 승률 간의 관계에 주목했는지를 이야기해 보겠습니다. 경기에서 이동은 단순히 이동 자체가 아닌 선수들의 피로도, 시차 적응, 그리고 환경 변화라는 세 가지 큰 영향을 미칩니다. 예를 들어, 장거리를 이동해야 할 때 선수들은 피로가 누적될 수밖에 없고, 멀리 갈수록 시차가 다르기 때문에 체력 회복이나 시차 적응에 더 많은 노력이 필요합니다. 이런 요인들이 경기에서 선수들의 퍼포먼스에 영향을 미칠 수 있습니다.

또한, 우리가 이 분석을 통해 얻고자 하는 건 전략적인 의사 결정 지원입니다. 분석 결과가 충분히 의미가 있다면, 앞으로 각 팀은 이동 일정이나 경기 전후의 훈련 계획 등을 좀 더 세밀하게 조정할 수 있게 되겠죠. 이로써 단기적인 승률뿐 아니라 팀의 전반적인 성적 관리와 운영에 도움이 될 수 있을 것으로 기대하고 있습니다.

슬라이드4

**데이터 수집**  
먼저, 이번 프로젝트에서 분석할 데이터를 수집하는 과정부터 설명드리겠습니다.  
저희는 **팀의 연도별 경기 스케줄** 데이터를 수집했는데, 이 과정에서 사용한 파일이 바로 baseball\_read\_2015\_2024\_short.py입니다. 이 파일은 지난 2015년부터 2024년까지의 모든 팀의 경기 일정을 불러오는 역할을 합니다. 이러한 스케줄 데이터를 통해 팀의 이동 거리나 승패와 같은 여러 변수를 도출할 수 있는 기초 데이터를 확보했습니다.

**데이터 정제**  
다음으로, 수집한 데이터를 분석에 적합하도록 정제했습니다. 데이터 정제 과정은 몇 가지 주요 작업으로 나누어지는데, 각각에 대해 설명드리겠습니다:

1. **구단 및 홈구장 이름 통일**  
   데이터에서 여러 구단이나 홈구장 이름이 서로 다르게 표기된 경우가 많았기 때문에, 이를 일관성 있게 통일하는 과정이 필요했습니다. 이 작업은 baseball\_read\_apply.py 파일을 사용하여 수행했습니다. 이 파일을 통해 구단 및 홈구장 명칭을 통일함으로써 이후 분석 단계에서 혼동 없이 데이터를 다룰 수 있게 되었습니다.
2. **이동거리 계산 및 누적거리 추가**  
   이어서 각 팀의 경기 이동 거리와 이를 누적한 거리를 계산했습니다. 이를 위해 baseball\_road\_apply.py 파일을 사용했으며, 이 파일은 각 팀이 경기마다 이동한 거리를 계산하고, 시즌 누적 거리 데이터를 생성하는 역할을 합니다. 이동거리가 누적되면, 시즌이 진행되면서 이동 피로가 누적되는 정도를 확인할 수 있어, 승률에 미치는 영향을 보다 심도 있게 분석할 수 있게 됩니다.
3. **승패 결과 및 누적 승률 추가**  
   마지막으로, 각 경기의 **승패 결과**와 **누적 승률**을 추가했습니다. 이를 위해 baseball\_winlose\_apply.py 파일을 사용했으며, 이 파일은 각 팀의 경기 결과를 기반으로 승패 데이터를 추가하고, 누적 승률을 계산하는 역할을 담당합니다. 이 데이터는 본격적인 승률 분석에 핵심적인 자료가 되며, 이동 거리와 승률 간의 상관관계를 파악하는 데 중요한 지표가 됩니다.

슬라이드5  
  
**분석 방법**  
프로젝트의 1차 목표는 각 팀의 이동거리가 경기 승률에 미치는 영향을 파악하는 것이었습니다. 이를 위해 **누적 이동거리**와 **승률의 변화를 시각화**하여 패턴을 찾고자 했습니다. 시각화 도구로는 Plotly Dash를 사용했습니다. 이 도구를 사용한 이유는 시각적 인터랙션을 통해 이동거리와 승률의 관계를 직관적으로 확인할 수 있기 때문입니다. 각 팀의 시즌별 누적 이동거리를 승률 변화와 함께 비교하면, 패턴이나 트렌드를 쉽게 포착할 수 있을 것이라고 예상했습니다.

**분석 결과**  
분석 결과, 이동거리와 승률 간에는 **유의미한 상관관계**를 발견하지 못했습니다. 기대했던 것과는 달리, 장거리 이동이 많은 경우에도 승률에 큰 변화가 나타나지 않거나, 일정한 패턴이 보이지 않는 팀이 많았습니다. 이동 거리가 팀의 성적에 미치는 영향이 예상보다 크지 않다는 결론에 도달하게 되었습니다.  
  
원인 분석  
**원인분석**  
저희가 이동거리와 승률의 상관관계를 좀 더 구체적으로 이해하기 위해 몇 가지 연도별 사례를 비교해 보았습니다.

1. **2021년 vs 2024년 (KIA의 승률 비교)**  
   2021년과 2024년의 KIA 팀 데이터를 살펴보면, **누적 이동거리**가 두 해 모두 비슷한 수준이었음에도 불구하고, **승률 차이가 매우 크게 나타났습니다.** 이로 인해 단순히 이동거리가 많다고 해서 팀 성적이 떨어지거나, 적다고 해서 성적이 향상되지 않는다는 점을 확인할 수 있었습니다. 즉, 이동거리 외에도 성적에 영향을 미치는 더 중요한 요인들이 존재할 가능성이 있습니다.
2. **2015년 vs 2022년 (전체 승률과 누적 이동거리 비교)**  
   반대로, 2015년과 2022년의 데이터를 보면, **누적 이동거리의 차이가 상당히 컸음에도 불구하고, 승률 차이는 미미했습니다.** 이 사례 역시 이동거리가 경기 승률에 미치는 영향이 크지 않다는 점을 보여줍니다. 따라서 이 데이터를 통해서도 이동거리가 팀 성적에 주요한 변수가 아니라고 해석할 수 있습니다.

**결론**  
위 사례들을 바탕으로, **이동거리가 팀의 경기 승률에 직접적인 영향을 미치지 않는다는 사실을 확인**했습니다. 분석 결과에 따르면, 여러 구단의 연도별 데이터를 통해 **평균 이동거리가 줄어드는 변화**를 확인할 수 있었으며, 이로 인해 **구단별 실제 이동거리 차이 또한 점차 줄어들고 있음을 알게 되었습니다.**  
이동거리가 점차 줄어드는 상황에서도 승률의 패턴이 크게 달라지지 않음을 보면, 앞으로는 이동거리 외에 승률에 영향을 줄 수 있는 다른 요인을 찾는 것이 중요하다는 결론을 내릴 수 있었습니다.

**결론**  
이 결과를 통해 **이동거리가 승률에 직접적인 영향을 미치지 않는다**는 사실을 확인할 수 있었습니다. 따라서, 원래 세웠던 1차 목표의 가설을 수정하거나, 새로운 분석 관점을 추가로 설정할 필요성이 생겼습니다. 앞으로는 이동거리 외에 **팀의 성적에 영향을 미칠 수 있는 다른 요인들**을 분석하는 방향으로 목표를 수정하여 프로젝트를 이어가고자 합니다.

슬라이드 6  
  
**목표 수정의 필요성**  
이전 분석을 통해, 이동거리가 경기 승률에 큰 영향을 미치지 않는다는 점을 확인했습니다. 따라서 **기존의 분석 목표**였던 이동거리와 승률의 관계를 넘어, 팀 성적에 영향을 줄 수 있는 **다른 요인들을 분석하는 방향으로 목표를 수정**하게 되었습니다. 이동거리가 성적에 미치는 영향이 크지 않다면, 그 외의 **다른 요소들이 성적에 더 중요한 역할**을 할 가능성이 높기 때문입니다.

**새로운 방향**  
새롭게 설정한 접근 방향은 **이동거리 외에 다른 변수들이 승률에 어떻게 영향을 미치는지 분석하는 것**입니다. 이를 위해 각 팀의 **타자, 투수, 수비**와 같은 세부 기록 데이터를 추가적으로 활용할 계획입니다. 예를 들어, 타자의 타격 성적, 투수의 방어율, 수비 지표 등이 팀 승률에 어떤 영향을 미치는지 분석할 것입니다. 이러한 변수들을 통해 팀 성적에 미치는 요인을 더 입체적으로 파악할 수 있을 것으로 기대하고 있습니다.

**목표 재설정**  
이와 같이 목표를 수정함에 따라, **팀 성적에 영향을 미치는 주요 요인**들을 파악하고, 이를 기반으로 **승률을 예측할 수 있는 모델을 개발**하는 것을 새로운 목표로 설정했습니다. 즉, 단순히 이동거리 외의 요인을 나열하는 것이 아니라, 성적과의 상관관계를 통계적, 혹은 머신러닝 모델을 통해 수치화하여 **팀이 최적의 성과를 낼 수 있는 전략적 데이터를 제공**하는 것을 목표로 하고 있습니다.

슬라이드 7   
**데이터 수집**  
이번 2차 목표에서는 팀 성적에 영향을 미칠 수 있는 다양한 요인을 분석하기 위해 보다 구체적인 데이터 수집이 필요했습니다. 이를 위해, 각 팀의 **타자, 투수, 수비 기록 데이터**를 수집했습니다. 이 데이터는 팀별 경기 결과뿐만 아니라, 선수들의 개별 성적을 통해 **어떤 요소가 승률에 중요한 역할을 하는지**를 파악할 수 있는 핵심 자료가 됩니다.  
또한, 팀 성적 변화의 큰 그림을 보기 위해 **연도별 승률 데이터**도 함께 수집했습니다. 이렇게 수집된 데이터를 바탕으로 시즌 성적에 대한 종합적인 분석이 가능해졌습니다.

**데이터 통합**  
수집한 데이터들을 활용해 **팀별 데이터를 하나의 데이터셋**으로 결합하는 과정이 필요했습니다. 이 작업에는 팀\_결합\_기록\_저장\_real.py 파일을 사용했습니다. 이 파일은 팀별로 나눠져 있는 타자, 투수, 수비 데이터를 **하나의 종합 데이터셋으로 통합**하는 역할을 수행하며, 이를 통해 팀 성적에 대한 다양한 요인을 한눈에 파악할 수 있게 준비했습니다.  
또한, 이렇게 통합된 데이터를 **숫자형 데이터로 변환**하여 본격적인 분석을 위한 전처리를 완료했습니다. 이 과정을 통해 각 지표가 분석에 적합하도록 가공되고, 예측 모델에 바로 사용할 수 있는 상태로 데이터를 준비했습니다.

슬라이드 8  
  
**데이터 정제**  
먼저, 데이터 분석에 앞서 **데이터 정제 작업**을 진행했습니다. 데이터를 살펴보니, 일부 결측값이 존재하여 이를 처리할 필요가 있었습니다. 결측값은 **평균값으로 대체**하여 데이터의 일관성을 유지하고, 분석 시 결과의 왜곡을 최소화하려 했습니다.  
또한, 수집한 데이터의 값들이 서로 다른 범위와 단위를 가지고 있어 **데이터 표준화**를 진행했습니다. 표준화를 통해 모든 데이터가 동일한 스케일에 맞춰졌으며, 이를 통해 분석의 정확성을 높이고, 특히 머신러닝 모델에 적합한 형태로 변환되었습니다.

**차원 축소  
다음 단계로 차원 축소 작업을 수행했습니다. 여기서 차원이란, 데이터에서 분석에 사용할 수 있는 변수나 특성의 수를 의미합니다. 예를 들어, 팀 성적을 예측할 때 팀의 연도별 경기 수, 승률, 주요 선수의 성적, 경기 일정 등 여러 가지 요소를 고려할 수 있습니다. 각각의 요소가 하나의 차원으로 이해될 수 있습니다.**

**데이터의 차원이 많아지면, 분석이 복잡해지고 모델이 오히려 성능이 떨어질 수 있습니다. 그래서 이 과정을 통해 데이터에서 핵심적인 정보만 남기고, 상대적으로 덜 중요한 정보는 제거하여 변수의 수를 줄이는 것이 차원 축소입니다. 이를 위해 \*\*주성분 분석(PCA)\*\*이라는 기법을 사용하여, 승률 예측에 중요한 특성들만 남겨 데이터의 차원을 줄였습니다. 이를 통해 분석이 보다 효율적으로 진행될 수 있도록 했습니다.**

**타겟 변수 및 피처 설정**  
이제 분석의 목표와 각 변수의 역할을 설정했습니다. **타겟 변수**는 팀의 **승률**로 설정하여, 궁극적으로 어떤 요소들이 승률에 영향을 주는지 예측하는 모델을 개발할 수 있도록 했습니다.  
**피처**로는 데이터에서 추출된 **숫자형 변수들**을 활용했습니다. 타격, 투수, 수비 등 각종 성적과 특성을 반영하는 변수들로서, 승률에 영향을 미칠 가능성이 높은 요소들을 다각적으로 분석할 수 있도록 구성했습니다.

슬라이드 9   
**사용한 회귀 모델**  
이번 분석에서는 다양한 회귀 모델을 사용해 팀의 승률을 예측해 보았습니다. 여러 회귀 모델을 적용한 이유는, **각 모델이 서로 다른 방식으로 데이터를 해석하고 예측**하기 때문입니다. 이를 통해 최적의 예측 성능을 가지는 모델을 찾고, 분석 결과의 신뢰성을 높이고자 했습니다.

구체적으로 사용한 모델은 다음과 같습니다:

* **선형 회귀 (Linear Regression)**  
  기본적인 선형 회귀 모델로, **각 변수와 승률 사이의 선형 관계**를 기반으로 예측합니다. 이 모델은 결과 해석이 직관적이라는 장점이 있습니다.
* **릿지 회귀 (Ridge Regression)**, **라쏘 회귀 (Lasso Regression)**, **엘라스틱넷 (ElasticNet)**  
  릿지와 라쏘 회귀는 각각 **규제를 통해 과적합을 방지**하고, 예측 모델의 안정성을 높입니다. 특히 라쏘 회귀는 중요하지 않은 변수를 자동으로 제외하는 효과가 있고, 엘라스틱넷은 릿지와 라쏘를 결합하여 두 방식의 장점을 살려 예측 성능을 강화합니다.
* **서포트 벡터 회귀 (SVR)**  
  비선형 데이터를 효과적으로 예측하기 위해 적용한 모델로, **데이터의 복잡한 관계**를 잘 반영할 수 있습니다.
* **결정 트리 회귀 (Decision Tree Regressor)**  
  결정 트리 모델은 데이터를 분할하여 **비선형 관계를 설명하는 데 적합**하며, 설명력이 높아 변수 간의 관계를 파악하는 데 유리합니다.
* **KNN 회귀 (K-Nearest Neighbors Regressor)**  
  이 모델은 **유사한 경기 상황을 기반으로 승률을 예측**하며, 데이터에 유사한 패턴이 있을 때 유용한 모델입니다.
* **랜덤 포레스트 회귀 (Random Forest Regressor)**  
  여러 결정 트리를 앙상블하여 예측 성능을 높이는 모델로, **변동성이 적고 예측의 안정성**이 높다는 장점이 있습니다.
* **그래디언트 부스팅 회귀 (Gradient Boosting Regressor)**  
  반복적인 학습을 통해 오차를 줄여가는 방식으로 예측 성능을 강화합니다. **복잡한 데이터 관계를 학습**하는 데 유용합니다.
* **XGBoost, LightGBM, CatBoost 회귀 모델**  
  이들 부스팅 모델은 **최적화 속도와 예측 성능이 매우 우수**해 많은 데이터와 복잡한 패턴이 있는 상황에서 뛰어난 성과를 보입니다. 특히, LightGBM과 CatBoost는 데이터의 분포와 변수 중요도를 고려하는 기능이 있어, 예측 성능을 높이는 데 도움이 됩니다.

슬라이드 11  
**데이터셋 구성**  
이번 분석에서는 **2015년부터 2023년까지**의 **10개 구단의 데이터**를 사용했습니다. 이 기간을 선택한 이유는, **최근 시즌의 성적 변화와 각 팀의 트렌드를 반영**하면서도 충분한 양의 데이터를 확보하기 위해서입니다. 또한, **2024년 이후의 데이터**는 **미래 예측 모델**을 위한 테스트 데이터로 **남겨두었으며**, 이 데이터를 통해 모델이 얼마나 정확하게 예측할 수 있는지 평가할 수 있습니다.

**모델 예측 결과**  
모델을 적용한 결과, 대부분의 팀에서 **실제 승률과 유사한 예측 승률**을 도출할 수 있었습니다. 예측과 실제 승률 간의 **오차 범위는 최소 0.001에서 최대 0.06**으로 나타났으며, 이는 모델이 각 팀의 승률을 **상당히 정밀하게 예측**할 수 있음을 보여줍니다.  
이 작은 오차 범위는 예측 모델이 **경기 결과 분석 및 예측에 신뢰성을 가지고 있음을 시사**하며, 각 팀의 성적과 관련된 전략적 의사결정에 활용될 수 있음을 의미합니다. 향후 2024년 데이터와의 비교를 통해 모델의 예측 정확도를 검증하고, 더 나아가 분석 결과를 팀 운영에 실질적으로 적용하는 방안도 고려하고 있습니다.

슬라이드 12  
**예측 결과 상세 분석**  
이번 슬라이드에서는 **각 팀별 예측 결과의 상세 정보**를 표로 정리했습니다. 실제 승률과 예측 승률을 비교하고, 각 팀에서 사용한 모델과 **PCA 구성요소 수**를 함께 표시하여, 예측 모델이 팀 성적을 얼마나 정확하게 반영하는지 설명드리겠습니다.

* **삼성: 예측 승률 0.497, 실제 승률 0.549로, 약간의 오차가 있지만 서포트 벡터 회귀(SVR) 모델을 통해 근사치를 예측했습니다. PCA 구성요소는 8개입니다.**
* **롯데: 예측과 실제 승률이 동일하게 0.471로, XGBoost Regressor 모델을 사용하여 매우 정확한 예측을 도출했습니다. 구성요소는 8개로 설정되었습니다.**
* **KIA: 예측 승률 0.614로 실제 승률 0.613에 거의 일치하는 값을 Ridge Regression 모델을 통해 예측했습니다. PCA 구성요소는 7개입니다.**
* **두산: 실제 승률 0.521에 대해 예측 승률이 0.528로 비슷한 결과를 얻었으며, Lasso Regression 모델을 사용했습니다. PCA 구성요소는 7개입니다.**
* **KT: 예측 승률 0.513과 실제 승률 0.507로, Linear Regression을 통해 상당히 근접한 예측을 했습니다. 8개의 PCA 구성요소를 사용했습니다.**
* **LG: 예측 승률 0.481, 실제 승률 0.535로 차이가 있었으며, KNN Regressor 모델을 통해 예측했습니다. 8개의 구성요소를 사용했습니다.**
* **SSG: 실제 승률 0.507에 대해 예측 승률은 0.515로 유사한 값을 도출했으며, Random Forest Regressor 모델을 사용했습니다. PCA 구성요소는 8개입니다.**
* **한화: 실제 승률 0.465, 예측 승률 0.443로 Lasso Regression 모델을 사용해 비교적 근사한 값을 예측했습니다. PCA 구성요소는 8개입니다.**
* **NC: 실제 승률 0.43에 대해 예측 승률은 0.483으로 차이가 있었지만, Lasso Regression 모델과 9개의 PCA 구성요소를 사용해 예측했습니다.**
* **키움: 예측 승률 0.405, 실제 승률 0.403으로, Lasso Regression을 사용해 거의 동일한 예측 결과를 도출했습니다. PCA 구성요소는 9개입니다.**

**결론**

**결론**  
전체적으로 보면, 대부분의 팀에서 실제 승률과 유사한 예측 승률이 도출되었으며, 특히 롯데와 KT, KIA 등은 예측과 실제 값의 차이가 거의 없었습니다. 이러한 결과는 각 팀의 성적 예측에서 모델이 높은 정확성을 보였다는 점을 시사합니다. PCA 구성요소 수를 조절하면서 모델 성능을 최적화했으며, 앞으로는 예측의 정확도를 높이기 위해 모델과 구성요소를 더욱 정교하게 조정할 계획입니다.

추가변인 입력  
  
**추가 변인 입력**  
이번 분석에서는 **연도별 예산 변화**와 **관중수 변화**를 추가 변수로 입력하여, 팀 성적에 대한 보다 정밀한 예측을 수행하고자 했습니다. 이 변수들을 포함한 이유는, 예산과 관중 수가 팀의 성적에 직간접적으로 중요한 영향을 미치기 때문입니다.

**예산 변화**  
연도별 팀 예산은 **팀 성적 향상에 직결되는 주요 요인 중 하나**로 볼 수 있습니다. 예산이 증가하면 팀은 더 뛰어난 선수 영입, 훈련 환경 개선, 스카우팅 활동 강화 등 성과에 중요한 투자를 할 수 있게 됩니다. 이러한 지원이 이루어지면 팀의 **전반적인 경기력이 향상될 가능성이 높아지며**, 이를 통해 시즌 성적이 긍정적으로 변화할 수 있습니다.  
특히, 예산 증가는 선수 영입뿐만 아니라 장기적으로 **팀의 조직력과 경쟁력**을 높이는 데도 큰 영향을 미칠 수 있습니다. 따라서 예산 변화를 변수로 추가함으로써, **팀 성적 예측의 정확도를 높이고, 예산과 성적 간의 관계를 분석**할 수 있습니다.

**관중수 변화**  
관중수는 팀의 성적과 상호작용하며 변화하는 경향이 있습니다. **팀 성적이 우수할수록 팬들의 관심**이 커지고, 이를 바탕으로 경기를 직접 관람하려는 관중의 수도 증가할 가능성이 큽니다. 반대로 팀의 성적이 부진할 경우, 팬들의 관심이 줄어들고 경기장 방문이 감소하여 **관중수가 줄어드는 현상**이 나타날 수 있습니다.  
이러한 관중수의 증감은 스포츠 전반에서 흔히 나타나는 현상으로, 팀의 흥행과 경기 운영에 중요한 요소입니다. 관중수 변화를 분석에 포함함으로써, 팀 성적이 **팬과의 관계에도 영향을 미친다는 사실을 반영**하고, 이를 통해 팀이 관중 흥행을 목표로 어떤 전략을 세울 수 있을지에 대한 시사점도 얻을 수 있습니다.

변인 추가 후 결과

 **삼성**: 예측 승률 0.497, 실제 승률 0.549로 약간의 차이가 있지만, **SVR 모델**을 사용하여 근사한 예측 결과를 얻었습니다. 8개의 PCA 구성요소를 사용했습니다.

 **롯데**: 예측 승률과 실제 승률이 0.471로 동일하게 나타났으며, **XGBoost Regressor** 모델을 사용해 높은 정확도를 보였습니다. PCA 구성요소는 8개로 설정되었습니다.

 **KIA**: 예측 승률과 실제 승률이 모두 0.613로, **Ridge Regression**을 통해 매우 근접한 예측을 도출했습니다. 구성요소 수는 7개입니다.

 **두산**: 실제 승률 0.521에 비해 예측 승률이 0.53으로 근접하게 나타났으며, **Lasso Regression** 모델을 사용했습니다. 7개의 PCA 구성요소를 사용했습니다.

 **KT**: 실제 승률이 0.507이고, 예측 승률은 0.502로, **Random Forest Regressor**를 통해 비교적 근접한 예측 결과를 얻었습니다. 구성요소 수는 7개입니다.

 **LG**: 예측 승률 0.524와 실제 승률 0.537로, **XGBoost Regressor** 모델을 사용하여 근사한 값을 예측했습니다. PCA 구성요소는 7개입니다.

 **SSG**: 예측 승률 0.502, 실제 승률 0.507로, **Lasso Regression**을 통해 근접한 예측을 얻었습니다. 8개의 구성요소를 사용했습니다.

 **한화**: 실제 승률 0.465에 비해 예측 승률이 0.443로, **Lasso Regression** 모델을 사용하여 약간의 오차가 있지만 유사한 예측값을 도출했습니다. 구성요소 수는 8개입니다.

 **NC**: 예측 승률이 0.483, 실제 승률이 0.43으로 차이가 있으나, **Lasso Regression** 모델과 9개의 PCA 구성요소를 사용하여 비교적 근접한 예측 결과를 얻었습니다.

 **키움**: 실제 승률이 0.403이고 예측 승률은 0.397로, **Ridge Regression** 모델을 사용하여 근사한 값을 예측했습니다. PCA 구성요소는 9개입니다.

참고사항

**예측과 실제 승률 차이의 원인 분석**  
삼성의 예측 승률과 실제 승률의 차이는 **홈구장의 타자친화적 특성**과 **특정 성적 지표의 변화**에서 기인했습니다. 특히, 2024 시즌 타격 지표를 분석한 결과, 몇 가지 주요 요소들이 확인되었습니다:

1. **타격 지표 변화**  
   올해 삼성의 타격 성적이 전년도 대비 개선되었으며, 특히 **홈런 수가 전년 대비 97개 증가**했습니다. 하지만 타격 지표가 전반적으로 상승한 것에 비해, **팀타율, 득점, 팀 안타** 수치는 여전히 리그 하위권에 머물렀습니다. 이는 홈런 수 증가에도 불구하고 팀의 타격 생산성이 리그 평균에 미치지 못했음을 의미합니다.
2. **홈구장의 타자친화적 특성**  
   삼성의 홈구장은 타자들에게 유리한 구장으로 알려져 있지만, 그럼에도 불구하고 **팀 홈런 수는 여전히 리그 하위권**에 속했습니다. 결과적으로 타자친화적 구장임에도 삼성은 많은 피홈런을 기록하며 “많이 맞고 적게 치는” 패턴을 보였으며, 이로 인해 승률이 예측과 다소 차이가 나는 원인으로 작용했습니다.
3. **투수진의 성과**  
   삼성의 투수진은 10년간 뚜렷한 지표 변화는 없었지만, **이번 시즌 전체 팀 투수 지표를 비교한 결과 리그 상위권**에 위치하고 있음을 확인했습니다.
   * **팀 방어율 3위**, **홀드 1위**, **세이브 3위**, **피안타 6위**, **볼넷 허용 9위**를 기록했으며, 특히 **WHIP(이닝당 출루 허용률)에서 리그 1위**를 차지했습니다.
   * 중간 계투의 **홀드 기록이 리그 최고 수준**으로 나타나, 계투진의 활약이 돋보였습니다. 또한 피안타와 볼넷 허용 수치가 적어, 전반적인 투수 성적이 우수했습니다.
4. **수비 지표**  
   삼성의 10년치 수비 데이터를 비교한 결과, 실책, 견제사, 병살 수치가 매우 낮았으며, 특히 **2024 시즌에는 리그 전체에서 실책(E)이 가장 적었습니다**. 그 결과, **FPCT(수비율)에서 리그 1위**를 기록하여 수비에서도 리그 최고 수준의 활약을 보였습니다. 10년 동안 지속적으로 개선된 실책, 견제사, 병살 지표는 삼성의 뛰어난 수비 역량을 나타내며, **리그 1위 수비력을 입증**하고 있습니다.

**종합 분석**  
삼성은 타자친화적 홈구장에서 많은 피홈런을 허용했음에도 불구하고, 리그 상위권의 투수 지표와 뛰어난 수비력을 바탕으로 승률을 유지할 수 있었습니다. 타격 지표의 변화와 홈런 수 증가는 팀 성적에 긍정적인 영향을 미쳤지만, 리그 전체 하위권의 타율 및 득점 생산성으로 인해 예측과 실제 성적 간에 차이가 발생한 것으로 보입니다.

**결론**

1. **이동거리와 승률**  
   첫 번째 목표였던 **이동거리와 승률의 관계**에서는 유의미한 상관관계를 발견하지 못했습니다. 즉, 팀의 이동거리가 승률에 직접적인 영향을 미치지 않는다는 결론에 도달했으며, 이 결과는 팀 성적에 영향을 미치는 더 중요한 요인들이 존재할 가능성을 시사했습니다.
2. **팀 성적 요인 분석**  
   이후 목표를 수정하여 팀의 **기록 데이터를 활용한 승률 예측**에 집중했습니다. 다양한 기록 데이터를 바탕으로 성적 예측을 수행한 결과, 팀 성적에 영향을 미치는 여러 요인을 성공적으로 분석할 수 있었고, **다양한 모델을 적용하여 최적의 예측 모델을 선정**하는 데도 성공했습니다. 이를 통해 각 팀의 성적을 더욱 정확하게 예측할 수 있었습니다.
3. **프로젝트 성과**  
   이번 프로젝트는 **데이터 분석의 전 과정을 경험**하고, 데이터를 수집하고 정제하는 기초 작업부터 고도화된 모델 적용과 예측 결과 분석까지 모두 다루는 중요한 경험이었습니다.  
   또한, 프로젝트 진행 중 여러 예측 모델을 실험하고, 목표를 유연하게 조정함으로써 **문제 해결 능력**과 **목표 설정의 중요성**을 깊이 인식하게 되었습니다. 이는 향후 데이터 분석 프로젝트를 수행하는 데 있어 매우 중요한 배움이었습니다.

**한계점 및 개선 방안**

1. **데이터의 양적 한계**  
   이번 분석에서는 **9년치의 데이터**를 사용했는데, 이는 데이터의 양이 상대적으로 적어 유의미한 장기적 패턴을 분석하기에 다소 부족할 수 있습니다. 특히, 팀 성적에 영향을 미치는 요소가 해마다 달라질 수 있기 때문에 **더 많은 연도와 다양한 팀의 데이터**를 추가할 필요가 있습니다. 이를 통해 보다 안정적이고 신뢰성 있는 예측 모델을 구축할 수 있을 것입니다.
2. **오차 원인 분석**  
   특정 팀에서는 예측 결과와 실제 성적 사이에 비교적 큰 오차가 발생하는 경우가 있었습니다. 이러한 **오차의 원인을 분석**하여 모델의 정확성을 개선하는 것이 필요합니다. 예를 들어, 팀별 고유의 경기 스타일이나 시즌 중의 예기치 못한 변수들이 오차에 영향을 미칠 수 있습니다. 앞으로는 팀별 특성을 반영하는 추가 데이터를 수집하고 모델을 조정함으로써 예측의 정밀도를 높이고자 합니다.
3. **추가 변수 고려**  
   현재 분석에는 정량적 변수들이 중심이 되었으나, **감독 교체, 선수 이적, 주요 선수의 부상**과 같은 비정량적 요인도 성적에 중요한 영향을 줄 수 있습니다. 이러한 요소들은 숫자로 나타내기 어렵지만, 앞으로는 팀 운영과 관련된 **비정량적 요인도 포함**하여 예측의 현실성과 정확도를 높이고자 합니다. 이를 통해 팀 성적을 더욱 입체적으로 예측할 수 있을 것입니다.