1. 승리와 거의 사실상 같은 변수 제거하기
: 모델이 너무 쉽게 '승리'를 예측하는 것을 방지하기 위해, 승리 결과와 매우 높은 상관관계를 가진 변수들을 제거하기.

- Rd (점수 차이): 승리와 0.9387의 상관관계를 가짐.
- Pitch_wins (투수 승리): 승리와 1.0000의 상관관계를 가짐. 사실상 승리와 같은 의미.
- L (패배): 승리와 거의 유사한 의미로 간주됨.

이러한 변수들은 '종속 변수(승리)'에 대한 정보를 너무 많이 담고 있어, 모델의 예측력이 과대평가되거나, 다른 변수들의 영향력을 제대로 분석하지 못하게 할 수 있음.

** '너무 쉽게 승리를 예측하는 것을 방지' 해야 하는 이유: 모델링에서는 최종 목적이 '단순한 예측'이 아니라, 승패에 영향을 미치는 핵심 요인(변수)을 이해하고 해석하기 위함.

- Rd나 Pitch_wins 처럼 승리 결과와 거의 동일한 정보를 가진 변수들을 사용하면 생기는 문제점.
1) 변수의 영향력 해석 불가능
>> 쉬운 예측의 문제: P_wins와 승리의 상관관계가 1.0000 --> 이 변수를 사용하면 모델은 "투수 승리 = 팀 승리"라는 사실을 학습하는 데 그치게 됨.
>> 목표 상실: 연구의 진짜 목적은 '툿 승리'가 아닌, '홈런 수', '탈삼진', '볼넷' 등의 경기 운영 요소가 승리에 얼마나 기여하는지 알아내는 것임. 승리와 직결되는 변수를 쓰면, 이 다른 변수들의 영향력이 가려지거나 왜곡되어 의미 있는 해석이 불가능해짐.

2) 모델의 일반화 능력 저하
>> 좋은 통계 모델은 '새로운 데이터'에서도 정확하게 예측할 수 있어야 함.
>> Rd(점수차이)는 이미 경기가 끝난 후의 결과(승패가 결정된 상황)를 반영하는 변수임. 이 변수를 상요해 승리를 예측한ㄴ 모델은 '원인'을 분석하는 것이 아니라, '결과'를 이용해 '결과'를 설명하는 꼴이 됨.
>> 이는 모델이 과거의 데이터를 과도하게 학습한 것으로 간주되어, 미래의 경기나 다른 리그에 적용할 때 예측 성능이 떨어질 수 있음.

"어떤 요소들이 승리를 가져오는가?"에 대한 통찰력을 얻는 것이 궁극적인 변수 선택 목적

2. 다중공선성 (Multicollinearity) 변수 제거하기

: 다중공선성은 독립 변수들 사이에 높은 상관관계가 존재할 때 발생하는 통계적 문제입니다. 이는 정보가 겹치는 변수가 있다는 것을 의미함.

다중공선성이 있으면 변수의 개별적인 영향력을 정확히 측정하기 어렵고, 모델의 안전성이 떨어질 수 있음.

- Ops (출루율+장타율): 이미 최종 변수 목록에 장타율 (slg)이 포함되어 있음. Ops는 slg를 포함하고 있어서 정보가 매우 많이 겹쳐서 다중공선성을 유발했을 가능성이 높음.
- R (팀 득점): 득점은 팀의 타격결과(total_hr, total_bb)와 타격 효율(slg)을 최종적으로 합산한 결과임. 이미 모델에 포함한 타격 관련 변수들의 '결과적인 지표'이기 때문에, 겹치는 정보 제거.
- avg_era (평균 자책점): 평균 자책점은 투수들의 허용한 시점(ra)을 경기 이닝으로 나눈 지표임. 최종 변수에 포함된 ra및 투구 효율(pitch_so, ptich_bb) 변수들과 정보가 겹쳐 제거.

3. 결론 및 최종 변수 후보 6개

: 다중공선성을 검토하는 지표인 "VIF (Variance Inflation Factor)" = 일반적으로 VIF 값이 10 이상이면 심각한 다중공선성이 있다면 판단하며, 선정된 6개 변수 후보들은 모두 VIF가 5.0 미만이어서 비교적 안정적인 변수들로 선택된 것으로 보임.

변수(variable)	VIF value	카테고리	설명
ra	1.15	수비/투구	상대에게 허용한 실점
slg	4.90	타격 효율	장타율
total_hr	3.01	타격 결과	팀 홈런 수
total_bb	2.21	타격/출루	팀 볼넷 수
pitch_so	2.47	투구 효율	투수 탈삼진 수
pitch_bb	2.30	투구 효율	투수 볼넷(피볼넷) 수

+ 선정된 6개 변수를 사용하여 승리를 예측하는 회귀 모델을 만들었을 때, 각 변수가 '승리에 미치는 영향력'을 어떻게 해석해야 할까?

해석: 다른 모델 변수(slg, totla_hr, ra 등)가 일정하다고 가정할 때, 해당 변수가 1단위 증가하면 승리 확률(또는 승리 수) 이 회귀 계수 값만큼 증가(+)하거나 감소(-)한다.

회귀 계수 방향: 실제 모델링 결과는 다를 수 있지만, 야구 상식을 바탕으로 각 변수가 승리 예측에 미치는 예상되는 영향의 방향을 해석 가능.

>> 승리에 긍정적인 영향 (Coefficient >0): 변수 계수가 양수가 나오면, 해당 지표가 높을수록 팀의 승리 확률이 증가한다는 의미.

>> 승리에 부정적인 영향 (Coefficient <0): 변수 계수가 음수로 나오면, 해당 지표가 높을수록 팀의 승리 확률이 감소한다는 의미.

**이때, 계수의 '절댓값 크기'가 변수의 승리에 미치는 상대적인 영향력을 나타냄!!

(다만, 이 변수들의 측정단위가 다르기 때문에, 실제 영향력 비교를 위해서는 표준화된 계수를 사용해야 함.) >> 방법 예시: 독립 변수를 Z-점수로 표준화하여 회귀 분석 실행 (변수값-평균/표준편차) or 파이썬이나 R에서는 출력 옵션에서 '표준화된 계수'를 선택해서 바로 결과를 얻을 수 있다고 함!