

세이버 스텟을 이용한  
한국 프로야구 선수의 연봉 데이터 분석



Baseball

# INDEX

상황분석

문제도출

데이터 분석

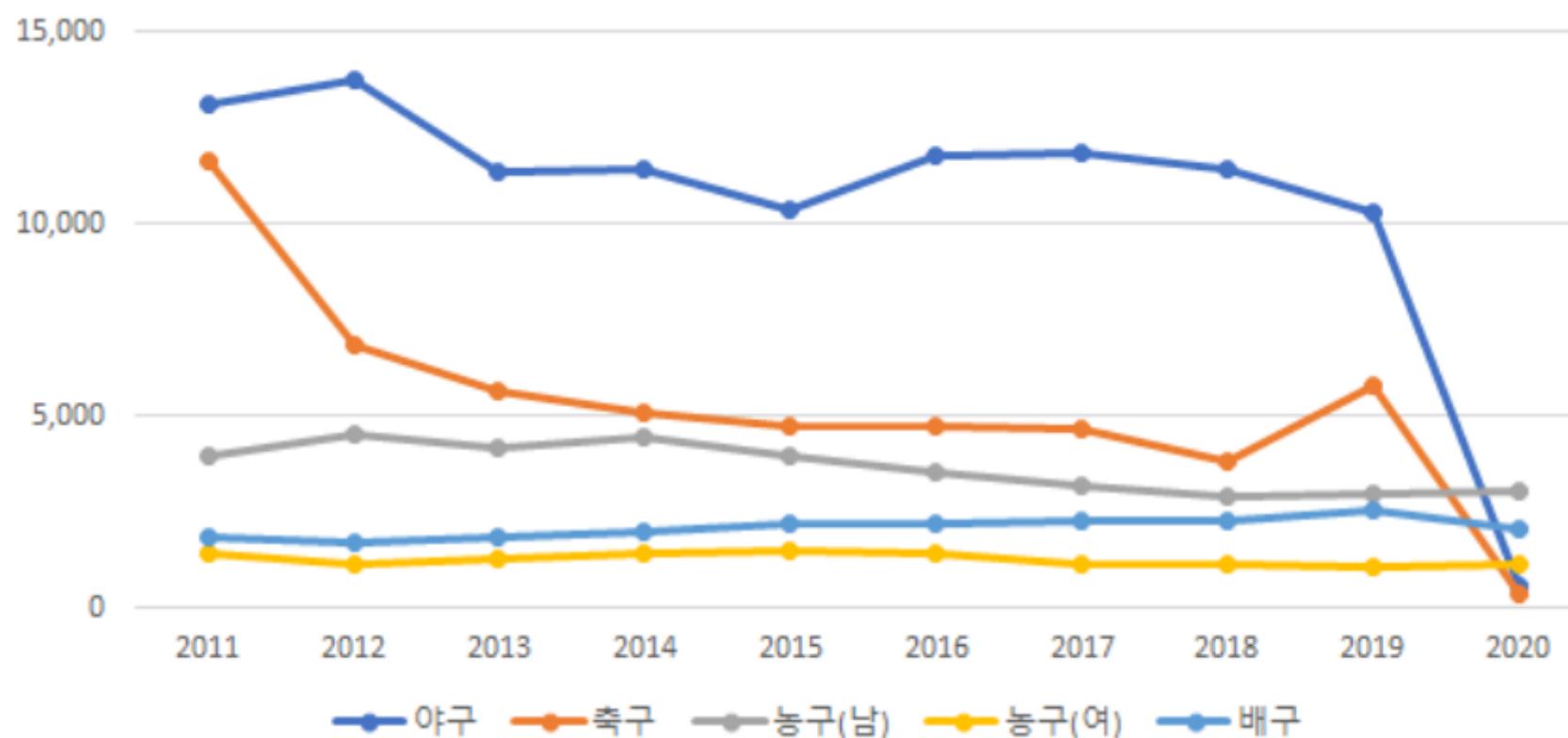
기대효과

- 1) 활용 및 메타 데이터 소개
- 2) 데이터 전처리 과정
- 3) 데이터 분석 - 타자
- 4) 데이터 분석 - 투수
- 5) 데이터 분석 결과

# 높은 국제 대회 성적과 메이저 리그에 진출하는 선수의 증가로 국내 스포츠 종목 중 프로야구는 압도적인 인기를 끌고있습니다.

## 압도적인 프로 야구 평균 관중수

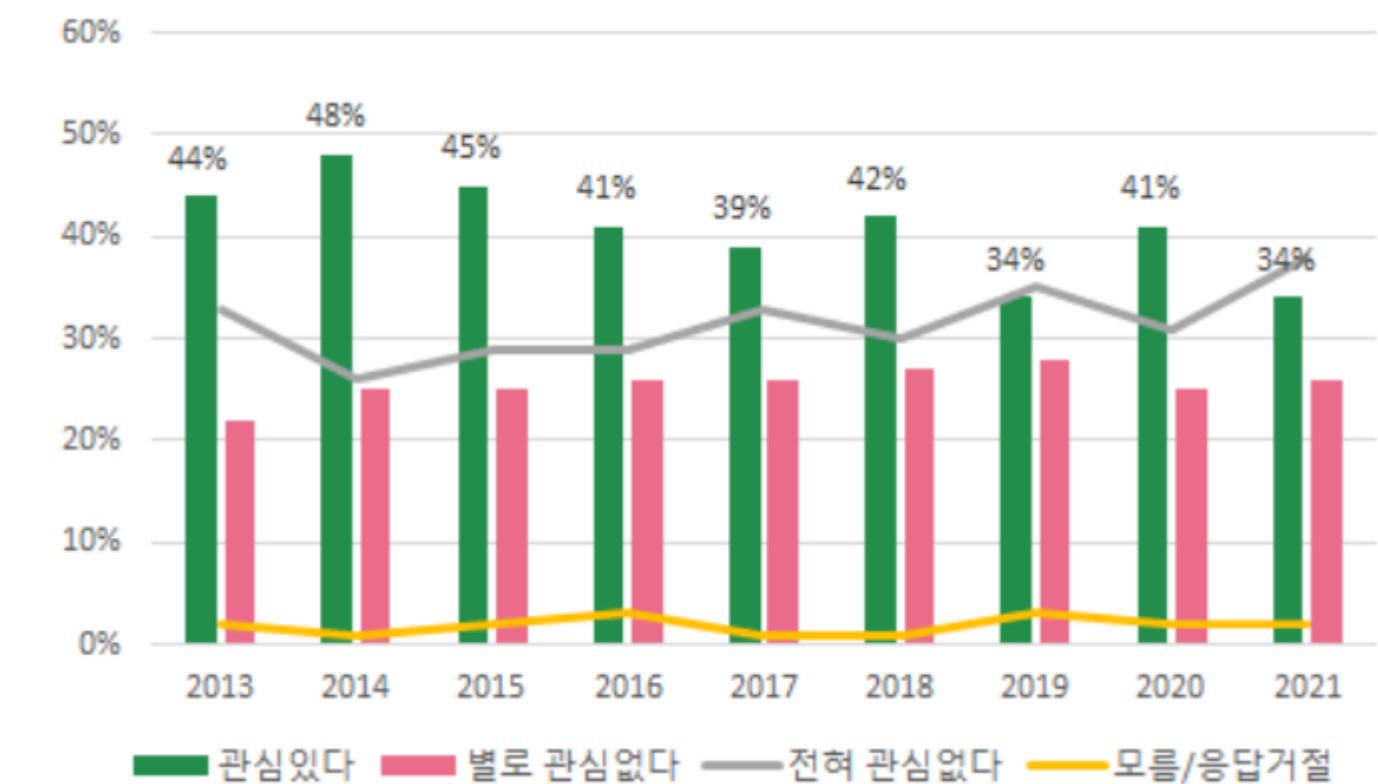
[주요 프로 스포츠 평균 관중 수]



출처: 문화체육관광부, 2021

## 프로야구에 대한 전 국민적 관심

[국내 프로야구 관심도]



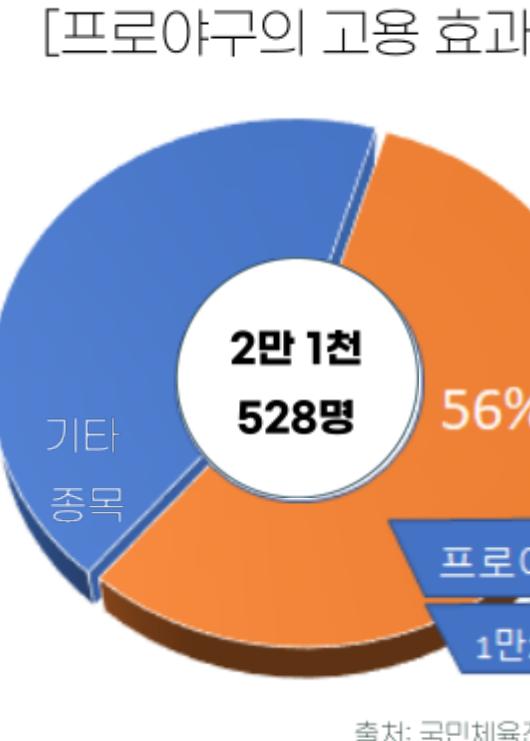
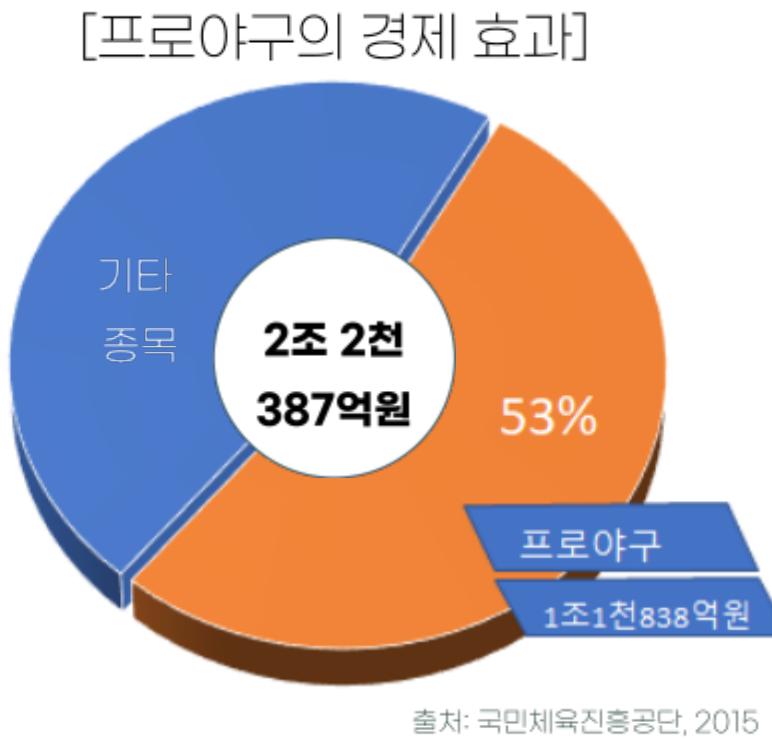
출처: 한국갤럽조사연구소, 2021

☞ 코로나 바이러스 감염증 발생 이전까지 프로야구는 9개년 간 국내 프로 스포츠 종목 중 가장 높은 평균 관중수를 기록했습니다.

☞ 매년 국민의 1/3 이상이 '프로야구에 대해 관심이 있다'고 응답할 정도로 프로야구는 국민의 대대적인 관심을 받는 스포츠입니다.

# 이러한 인기에 힘입어 프로야구는 브랜드 마케팅 효과와 상권 활성화 등 경제 및 고용 파급효과도 확대되고 있습니다.

## 프로야구의 높은 경제적 파급 효과



## 프로야구의 경제적 가치

[프로야구단 부분별 가치평가 순위] (단위 : 억원)

| 순위 | 구단          | 시장 가치 | 경기장 가치 | 스포츠가치 |     |     | 총액   |
|----|-------------|-------|--------|-------|-----|-----|------|
|    |             |       |        | 연봉    | 중계  | 성적  |      |
| 1  | 두산 BEARS    | 370   | 1099   | 117   | 166 | 155 | 1907 |
| 2  | LG TWINS    | 370   | 1139   | 112   | 167 | 95  | 1883 |
| 3  | SK WARRIORS | 340   | 788    | 127   | 166 | 125 | 1546 |
| 4  | 롯데 GIANTS   | 394   | 650    | 131   | 160 | 85  | 1420 |
| 5  | 삼성 LIONS    | 281   | 686    | 97    | 161 | 150 | 1375 |
| 6  | 키움 DRAGONS  | 370   | 616    | 75    | 175 | 82  | 1318 |
| 7  | NC DOLPHINS | 120   | 748    | 104   | 166 | 66  | 1204 |
| 8  | KIA TIGERS  | 168   | 519    | 119   | 211 | 180 | 1197 |
| 9  | 한화 EAGLES   | 170   | 554    | 100   | 214 | 84  | 1122 |
| 10 | KT WILDCATS | 140   | 361    | 81    | 163 | 67  | 812  |

출처: 통계청과 KBO, 2019

☞ 배구, 농구, 야구, 축구의 국내 4대 프로리그 중  
프로 야구의 경제 · 고용 효과가 과반 이상을 차지합니다.

☞ 프로야구 각 구단의 경제적 가치는  
총 1조 4000억 원에 달할 정도입니다.



[ 프로야구 데이터를 활용한 게임 ]

특히, 야구는 수치화하기 쉬운 평가 지표를 가지기 때문에 데이터 기반의 '세이버 매트릭스'가 보편화되고 있습니다.

☞ 사회과학의 게임 이론과 통계학적 방법론을 도입하여 야구의 객관적인 지식을 추구하는 세이버 매트릭스(SABERmetrics)가 빅데이터의 발전에 힘입어 보편화되고 있습니다.

☞ 프로야구 데이터를 활용하여 '아웃 오브 더 파크 베이스볼', '컴투스 프로야구 2021', '프로야구 매니저', '마구마구' 등의 다수의 PC와 모바일 게임이 출시되었습니다.

# 하지만, 기존의 야구 데이터 분석은 클래식 스탯을 사용하거나 팀 관련 데이터만을 활용하여 승·패만을 예측하는 한계가 존재합니다.

## 오류가 많고 예측이 정확하지 않은 클래식 스탯

KBO리그기록 및 순위 [주요 프로 스포츠 평균 관중 수]

2008.

팀순위

| 순위 | 팀   | 경기수 | 승  | 패  | 무 | 승률    | 개인차  | 연속 | 출루율   | 장타율   | 최근10경기   |
|----|-----|-----|----|----|---|-------|------|----|-------|-------|----------|
| 1  | SK  | 126 | 83 | 43 | 0 | 0.659 | 0.0  | 2패 | 0.361 | 0.404 | 5승-5패-0무 |
| 2  | 두산  | 126 | 70 | 56 | 0 | 0.556 | 13.0 | 2패 | 0.354 | 0.378 | 3승-7패-0무 |
| 3  | 롯데  | 126 | 69 | 57 | 0 | 0.548 | 14.0 | 1패 | 0.356 | 0.405 | 4승-6패-0무 |
| 4  | 삼성  | 126 | 65 | 61 | 0 | 0.516 | 18.0 | 1승 | 0.344 | 0.376 | 5승-5패-0무 |
| 5  | 한화  | 126 | 64 | 62 | 0 | 0.508 | 19.0 | 2승 | 0.333 | 0.395 | 6승-4패-0무 |
| 6  | KIA | 126 | 57 | 69 | 0 | 0.452 | 26.0 | 1승 | 0.336 | 0.352 | 5승-5패-0무 |
| 7  | 넥센  | 126 | 50 | 76 | 0 | 0.397 | 33.0 | 1승 | 0.331 | 0.368 | 6승-4패-0무 |
| 8  | LG  | 126 | 46 | 80 | 0 | 0.365 | 37.0 | 1승 | 0.321 | 0.353 | 4승-6패-0무 |

출처 : KBO 기록실

☞ 출루율, 장타율 등의 클래식 스탯(Classic Stats)은 이해하기 쉽고 가독성이 높으나 오류가 많고 예측이 정확하지 않다는 한계가 존재합니다.

## 팀 관련 데이터를 사용한 기준 승·패예측 데이터 분석

데이터마이닝을 활용한 한국프로야구 승패예측모형  
수립에 관한 연구

오윤학 · 김 한 · 윤재섭 · 이종석<sup>†</sup>

성균관대학교 시스템경영공학과

☞ 로지스틱 회귀 분석과 의사결정나무모형(CHAID 기법) 등을 사용하여 프로야구의 승·패 예측모형에 관한 연구 수행

혼합형 기계 학습 모델을 이용한 프로야구  
승패 예측 시스템

(Win/Lose Prediction System : Predicting Baseball Game  
Results using a Hybrid Machine Learning Model)

홍석미 \* 정경숙 \*\* 정태충 \*\*\*  
(SeokMi Hong) (KyungSook Jung) (TaeChong Chung)

☞ ID3, 통계적 방법, 역전파 알고리즘을 사용하여 승패 예측 시스템을 구축하는 연구 수행

## 또한, 단순한 팀의 승·패 예측에서 나아가 데이터 분석을 토대로 한 연봉 제안이 시급한 실정입니다.

**FA 먹튀, 연봉 거품이라는 용어가 등장할 정도로 선수 연봉에 대한 의문 급증**

### [김수인의 직격 야구] FA(자유계약선수) 먹튀, 잘 살피자

그렇지만 FA가 부메랑으로 돌아와 감독의 뒤통수를 치는 일이 더러 있다. 대표적인 팀이 롯데다. 롯데는 개막 5연승의 신바람을 내며 올시즌 포스트시즌 진출의 희망가를 불렀지만 27일 현재 24 일째 8위로 제자리 걸음, 힘겨운 5강 싸움을 벌이고 있다(5위 LG에 3.5게임차).

출처 : 스포츠 한국, 2020

☞ 자유계약선수(FA)에 높은 연봉을 제시하여 영입하였으나 팀의 승리에는 도움이 되지 않는 경우가 빈번했습니다.

## 거품 빼자더니, FA 대어들에겐 '한파' 없었다

빅4의 계약이 주는 의미는 결과적으로 대어급 선수들에게는 'FA 한파란 딴 세상 이야기'였음이 증명되었다는 사실이다. 최근 몇 년간 FA 시장의 거품에 대한 자성 여론과 코로나19로 인한 모기업 투자 축소- 프로구단들의 젊은 선수 육성 기조 등으로 인하여 올해의 FA들은 과거 만큼의 대박을 기대하기 어려울 것이라는 전망도 있었다. 하지만 막상 뚜껑을 열자 상황이 달라졌다.

출처 : 오마이뉴스, 2020

☞ 자유계약선수(FA) 연봉에 대한 자성여론에도 연봉 거품은 최근까지도 계속되고 있습니다.

**조정을 신청할 정도로 합의점을 찾기 어려운 연봉 협상**

### '초미의 관심' kt 주권 연봉 칼자루 뿐 인사 결정

주권은 올해 연봉으로 지난해보다 1억 원 오른 2억5000만 원을 주장했지만 kt는 2억2000만 원을 고수해 협상이 결렬됐다. 지난해 77경기 70이닝을 던진 주권은 6승 2패 31홀드 평균자책점 2.70의 성적으로 홀드왕에 올랐고, kt의 창단 첫 포스트시즌 진출에 기여했다.

역대 KBO 연봉 조정위는 20번 열렸는데 선수가 이긴 것은 단 한번뿐이었다. 2002년 LG 내야수였던 류지현 현 LG 감독이다. 당시 류 감독은 1000만 원 삭감된 연봉 1억9000만 원을 부른 LG와 맞서 2억2000만 원을 요구해 조정위에서 이겼다.

선수가 이길 확률은 5%에 불과한 상황. 다만 최근 달라진 프로야구계의 분위기와 팬들의 여론 압박 등 여러 조건들을 보면 KBO 연봉 조정위가 구단에만 유리하게 결정을 내리기 쉽지 않을 것이라는 전망이 나오고 있다.

출처 : 노컷뉴스, 2021

☞ KBO에 조정을 맡기는 연봉 조정 신청이 발생할 정도로 선수와 구단이 연봉에 대한 의견 차를 좁히지 못하는 경우가 있습니다.

# 야구는 각 포지션이 팀 승·패에 영향을 미치기 때문에 타자와 투수 개인의 역량을 정확하게 분석하는 것이 중요합니다.

## 선수 영입으로 우승을 거머쥔 NC

### 양의지, 공통의 심장을 뛰게 하다...NC, '최고포수' 영입 효과톡톡

#### 창단 후 첫 꼴찌 기록했던 NC, 양의지 영입으로 반등

초대 사령탑이었던 김경문 감독의 시즌 중 사퇴 등의 충격파로 팀이 최하위로 미끄러진 터라 터닝포인트가 필요했던 NC였다. KBO리그 9번째 심장을 자처하던 NC가 리그 꼴찌로 추락한 것은 2011년 8월 팀 창단 이후 처음 있는 일. NC는 2013년 7위(당시 9개 팀)로 1군 리그 데뷔 신고식을 치른 뒤 2014년부터 2017년까지 4년 연속 가을야구에 진출했다. 시즌 직후 계약한 이동욱 신임 감독 체제에서 분위기 전환이 급선무였다.

더불어 NC는 2019년 새 홈구장인 창원 NC파크 개장을 앞두고 있었다. 경남과 창원이 1200억원을 들여 만든 메이저리그 식 구장이다. 프로야구 출범 팀인 롯데 팬들이 경남과 창원 지역 곳곳에 포진한 상황에서 NC는 흥행의 새로운 동력이 필요했고 양의지는 새 야구장에 새로운 숨을 불어넣을 최고의 스타였다.

NC는 양의지에게 4년 총액 125억원(계약금 60억원+4년 연봉 65억원)의 거액을 제시했다. NC와 뒤늦게 '쩐의 전쟁'을 벌이게 된 두산은 부랴부랴 액수를 키웠으나 양의지를 붙들지는 못했다. 125억원은 일본과 미국 무대를 거쳐 롯데로 돌아온 이대호(150억원)에 이어 역대 두 번째로 많은 FA 계약액이었다. 게다가 플러스-마이너스 옵션이 전혀 없는 완전 보장 액수였다. NC가 얼마나 간절히 양의지를 원했는지 알 수 있는 대목이다. 양의지와의 협상 테이블에 앉았던 김종문 NC 단장은 "막바지에는 두산도 얼추 비슷한 액수를 제시한 것으로 안다. 돈도 돈이지만 양의지가 또 다른 환경에서 자신의 가치를 증명하고 싶었던 것 같다. 협상 과정에서 그런 점을 느꼈다"고 했다.

출처: 시사저널, 2020

성공적인 연봉협상을 통해 양의지 개인은 리그 타격왕<sup>(0.354)</sup>에 올랐고 장타율<sup>(0.574)</sup>, 출루율<sup>(0.438)</sup>도 1위를 기록했으며, 팀을 우승으로 이끌었습니다.

## 클래식 스탠보다 정확도가 높은 세이버 스탠



| 순  | 이름    | 팀       | 생산력+<br>wRC+ | 타석  | HR%  | BB%  | K%   | BB/K | IsoP | IsoD | BABIP | Spd | PSN   | 타격 생산력 |       |                   |      |
|----|-------|---------|--------------|-----|------|------|------|------|------|------|-------|-----|-------|--------|-------|-------------------|------|
|    |       |         |              |     |      |      |      |      |      |      |       |     |       | wOBA   | wRC   | wRC <sub>27</sub> | wRAA |
| 1  | 로하스   | 20 K RF | 180.8        | 628 | 7.48 | 10.4 | 21.0 | 0.49 | .331 | .068 | .383  | 2.6 | 0.00  | .467   | 149.5 | 10.68             | 67.0 |
| 2  | 최형우   | 20 K DH | 168.4        | 600 | 4.67 | 11.7 | 16.8 | 0.69 | .236 | .079 | .397  | 2.5 | 0.00  | .450   | 134.1 | 10.37             | 55.3 |
| 3  | 라모스   | 20 L 1B | 153.8        | 494 | 7.69 | 11.1 | 27.5 | 0.40 | .313 | .084 | .314  | 3.7 | 3.80  | .407   | 91.8  | 7.70              | 27.0 |
| 4  | 양의지   | 20 N C  | 153.3        | 528 | 6.25 | 8.7  | 8.9  | 0.98 | .276 | .073 | .305  | 3.5 | 8.68  | .432   | 109.6 | 8.97              | 40.3 |
| 5  | 나성범   | 20 N DH | 152.2        | 584 | 5.82 | 8.4  | 25.3 | 0.33 | .272 | .067 | .395  | 4.6 | 5.51  | .430   | 120.4 | 8.96              | 43.8 |
| 6  | 강백호   | 20 K 1B | 149.3        | 574 | 4.01 | 11.5 | 16.2 | 0.71 | .214 | .081 | .367  | 4.3 | 10.73 | .421   | 113.6 | 8.84              | 38.3 |
| 7  | 터커    | 20 K RF | 148.9        | 631 | 5.07 | 12.0 | 10.6 | 1.13 | .251 | .092 | .300  | 2.3 | 0.00  | .420   | 124.2 | 8.51              | 41.4 |
| 8  | 김현수   | 20 L LF | 148.4        | 619 | 3.55 | 10.2 | 8.6  | 1.19 | .192 | .067 | .332  | 3.2 | 0.00  | .407   | 115.2 | 8.10              | 33.9 |
| 9  | 김하성   | 20 키 SS | 147.4        | 622 | 4.82 | 12.1 | 10.9 | 1.10 | .218 | .092 | .304  | 5.2 | 26.04 | .407   | 115.9 | 8.03              | 34.3 |
| 10 | 페르난데스 | 20 두 DH | 146.5        | 668 | 3.14 | 8.7  | 6.3  | 1.38 | .157 | .065 | .333  | 1.7 | 0.00  | .406   | 123.6 | 7.85              | 35.9 |

출처: STATIZ

☞ 타율·홈런·타점 등 고전적인 스탠(Classic Stats)을 과학적 장비 및 계산식을 적용한 세이버 스탠(Saber Stats)을 사용하면 선수들의 능력치를 자세하고 정확하게 분석 가능합니다.

# 야구는 각 포지션이 팀 승·패에 영향을 미치기 때문에 타자와 투수 개인의 역량을 정확하게 분석하는 것이 중요합니다.

선수 영입으로 우승을 거머쥔 NC

양의지, 공통의 심장을 뛰게 하다...NC, '최고 포수' 영입 효과톡톡

창단 후 첫 끌찌 기록했던 NC, 양의지 영입으로 우승을 거머쥔 NC

초대 사령탑이었던 김경문 감독의 시즌 중 사퇴 등의 충격파로 팀이 최하위로 미끄러진 터라 터닝포인트가 필요했던 NC였다. KBO리그 9번째 심장을 자처하던 NC가 리그 끌찌로 추락한 것은 2011년 8월 24일 이후 처음이다. 2013년 7위(당시 9개 팀)로 1군 리그 데뷔 신고식을 치른 뒤 2014년부터 2017년까지 4년 연속 9위에 머물렀다. 그 즈음 직후 계약한 이동욱 신임 감독 체제에서 분위기 전환이 급선무였다.

더불어 NC는 2019년 새 홈구장인 창원 NC파크 개장을 앞두고 있었다. 경기장은 물론 구장 구조, 관중 편의시설 등 구장이다. 프로야구 출범 팀인 롯데 팬들이 경남과 창원 지역 곳곳에 포진해 있어 NC는 새 홈구장에 대한 기대감이 커졌다. 특히 양의지는 새 야구장에 새로운 숨을 불어넣을 최고의 스타였다.

NC는 양의지에게 4년 총액 125억원(계약금 60억원+4년 연봉 65억원)의 거액을 제시했다. NC와 뒤늦게 '쩐의 전쟁'을 벌이게 된 두산은 부랴부랴 액수를 키웠으나 양의지를 볼들지는 못했다. 125억원은 일본과 미국 무대를 거쳐 롯데로 돌아온 이대호(150억원)에 이어 역대 두 번째로 많은 FA 계약액이었다. 게다가 플러스-마이너스 옵션이 전혀 없는 완전 보장 액수였다. NC가 얼마나 간절히 양의지를 원했는지 알 수 있는 대목이다. 양의지와의 협상 테이블에 앉았던 김종문 NC 단장은 "막바지에는 두산도 얼추 비슷한 액수를 제시한 것으로 안다. 돈도 돈이지만 양의지가 또 다른 환경에서 자신의 가치를 증명하고 싶었던 것 같다. 협상 과정에서 그런 점을 느꼈다"고 했다.

출처: 시사저널, 2020

클래식 스탠보다 정확도가 높은 세이버 스탠

STATIZ

**따라서, 세이버 스탠을 사용하여  
선수 개인의 역량을 분석하여  
연봉을 예측하는 것이 필요합니다.**

| 순위 | 선수명   | 포지션  | 경기수   | 득점  | 득점률  | 타율   | OPS   | BB/K | IsoP | IsoD | BABIP | Spd | PSN   | 타격 생산력 |       |                   |      |
|----|-------|------|-------|-----|------|------|-------|------|------|------|-------|-----|-------|--------|-------|-------------------|------|
|    |       |      |       |     |      |      |       |      |      |      |       |     |       | wOBA   | wRC   | wRC <sub>27</sub> | wRAA |
| 1  | 로하스   | K RF | 180.8 | 628 | 7.48 | 10.4 | .21.0 | 0.49 | .331 | .068 | .383  | 2.6 | 0.00  | .467   | 149.5 | 10.68             | 67.0 |
| 2  | 김현수   | L LF | 153.8 | 494 | 7.69 | 11.1 | .27.5 | 0.40 | .313 | .084 | .314  | 3.7 | 3.80  | .450   | 134.1 | 10.37             | 55.3 |
| 3  | 강백호   | K 1B | 149.3 | 574 | 4.01 | 11.5 | .16.2 | 0.71 | .214 | .081 | .367  | 4.3 | 10.73 | .421   | 113.6 | 8.84              | 38.3 |
| 4  | 터커    | K RF | 148.9 | 631 | 5.07 | 12.0 | .10.6 | 1.13 | .251 | .092 | .300  | 2.3 | 0.00  | .420   | 124.2 | 8.51              | 41.4 |
| 5  | 김하성   | K SS | 147.4 | 622 | 4.82 | 12.1 | .10.9 | 1.10 | .218 | .092 | .304  | 5.2 | 26.04 | .407   | 115.9 | 8.03              | 34.3 |
| 6  | 페르난데스 | DH   | 146.5 | 668 | 3.14 | 8.7  | .6.3  | 1.38 | .157 | .065 | .333  | 1.7 | 0.00  | .406   | 123.6 | 7.85              | 35.9 |

출처: STATIZ

성공적인 연봉협상을 통해 양의지 개인은 리그 타격왕<sup>(0.354)</sup>에 올랐고  
장타율<sup>(0.574)</sup>, 출루율<sup>(0.438)</sup>도 1위를 기록했으며, 팀을 우승으로 이끌었습니다.

☞ 타율·홈런·타점 등 고전적인 스탠(Classic Stats)을 과학적 장비 및  
계산식을 적용한 세이버 스탠(Saber Stats)을 사용하면  
선수들의 능력치를 자세하고 정확하게 분석 가능합니다.

## 데이터 분석 절차

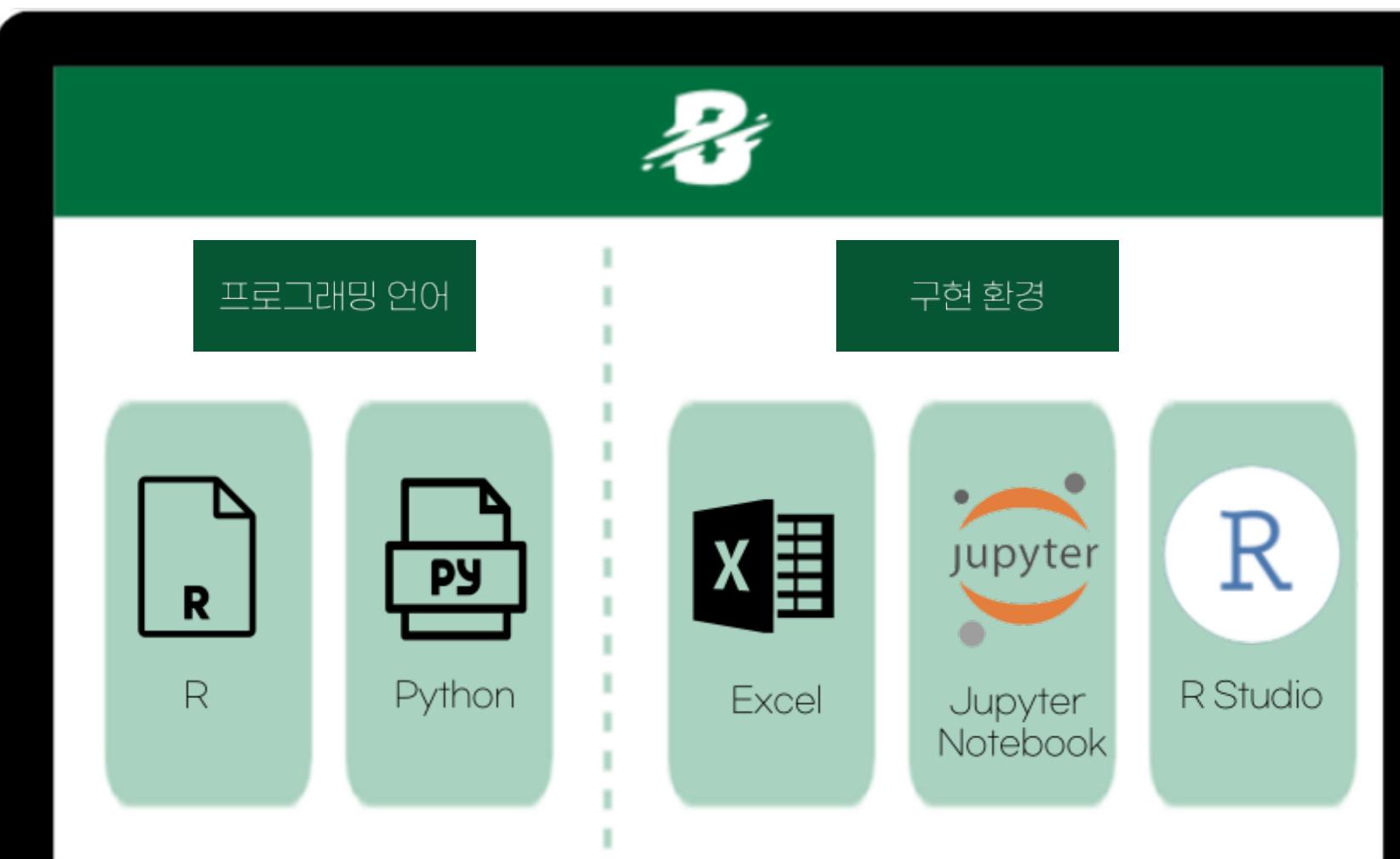
- 1) 활용 및 메타 데이터 소개  
어떤 데이터를 수집하고,  
어떤 도구를 이용하여 데이터를 분석했는지 설명
  - 2) 데이터 전처리 과정  
데이터 분석에 있어서 누락된 항목,  
분석이 필요 없는 항목을 제거하는 과정  
(중복값, 결측값, 이상치 등을 처리)
  - 3) 데이터 분석 - 타자  
회귀분석 모형을 사용하여  
타자의 세이버 스탯 및 기존 연봉 정보로  
미래 연봉 예측 모델 설계
  - 4) 데이터 분석 - 투수  
회귀분석 모형을 사용하여  
투수의 세이버 스탯 및 기존 연봉 정보로  
미래 연봉 예측 모델 설계
  - 5) 데이터 분석 결과  
분석 데이터의 패턴 도출과  
인사이트 발견

## 간트 차트(Gantt Chart)

☞ 프로젝트 일정을 간트 차트(Gantt Chart)를 사용하여 관리

# 프로그래밍 언어 및 구현 환경

- ☞ 웹 스크래핑을 통해 데이터를 수집한 뒤 R을 사용하여 데이터를 전처리하고, Python 을 사용하여 데이터 분석 모델을 구축



# 활용 데이터 수집

The screenshot shows a search interface for KBO statistics. The results table displays player statistics for the 2020 season, including fields like 선수명 (Player Name), 팀 (Team), and various batting and pitching metrics.

| 순위 | 선수명    | 팀   | Avg   | G   | PA  | AB  | R   | H   | 2B | 3B | HR | TB  | RBI | SAC | SF |
|----|--------|-----|-------|-----|-----|-----|-----|-----|----|----|----|-----|-----|-----|----|
| 1  | 최철우    | KIA | 0.354 | 140 | 600 | 522 | 93  | 185 | 37 | 1  | 28 | 308 | 115 | 0   | 3  |
| 2  | 손아섭    | 롯데  | 0.352 | 141 | 611 | 540 | 98  | 190 | 43 | 0  | 11 | 266 | 85  | 1   | 7  |
| 3  | 로하스    | KT  | 0.349 | 142 | 628 | 550 | 116 | 192 | 39 | 1  | 47 | 374 | 135 | 0   | 8  |
| 4  | 박민우    | NC  | 0.345 | 126 | 530 | 467 | 82  | 161 | 27 | 5  | 8  | 222 | 63  | 2   | 10 |
| 5  | 페르나리에스 | 토신  | 0.340 | 144 | 668 | 586 | 104 | 199 | 29 | 0  | 21 | 291 | 105 | 0   | 11 |

출처 : KBO

## ① KBO 기록실

☞ 기간 : 2018년도 ~ 2020년도

☞ 데이터 분류 : 타자, 투수

☞ 시즌 : KBO 정규 시즌

☞ 활용변수 : 클래식 스탯 데이터

1) 타자 : H(안타), HR(홈런), RBI(타점)

2) 투수 : ERA(평균자책점), IP(이닝), WHIP(이닝당 출루허용률)

The screenshot shows a search interface for STATIZ statistics. The results table displays player statistics for the 2020 season, including fields like 이름 (Name), 팀 (Team), and various offensive and defensive metrics.

| 순 | 이름  | 팀       | 생산력<br>wRC+ | 타석  | HR%  | K%   | BB/K | IsoP | IsoD | BABIP | Spd  | PSN | 타격 생산력 | wOBA | wRC   | wRC27 | wRAA | wRC+ |       |       |      |       |
|---|-----|---------|-------------|-----|------|------|------|------|------|-------|------|-----|--------|------|-------|-------|------|------|-------|-------|------|-------|
| 1 | 로하스 | 2B 4 SF | 180.8       | 628 | 7.48 | 10.4 | 21.0 | .49  | .331 | .068  | .383 | 2.6 | 0.00   | .467 | 149.5 | 10.68 | 67.0 | .467 | 149.1 | 10.65 | 66.7 | 180.8 |
| 2 | 최철우 | 2B 4 SF | 168.4       | 600 | 4.87 | 11.7 | 16.8 | .69  | .236 | .079  | .397 | 2.5 | 0.00   | .450 | 134.1 | 10.37 | 55.3 | .448 | 132.7 | 10.28 | 53.9 | 168.4 |
| 3 | 라모스 | 2B 1 SF | 153.8       | 494 | 7.09 | 11.1 | 27.5 | .40  | .313 | .084  | .314 | 3.7 | 3.80   | .407 | 91.8  | 7.70  | 27.0 | .425 | 99.7  | 8.38  | 34.9 | 153.8 |
| 4 | 양의지 | 2B 9 C  | 153.3       | 528 | 6.25 | 8.7  | 8.9  | .58  | .270 | .073  | .305 | 3.5 | 8.68   | .432 | 109.6 | 8.97  | 40.3 | .425 | 106.3 | 8.70  | 37.0 | 153.3 |
| 5 | 나성범 | 2B 9 SF | 152.2       | 584 | 5.82 | 8.4  | 25.3 | .33  | .272 | .067  | .395 | 4.6 | 5.51   | .430 | 120.4 | 8.96  | 43.8 | .423 | 116.7 | 8.68  | 40.0 | 152.2 |

출처 : STATIZ

## ② STATIZ 기록실

☞ 기간 : 2018년도 ~ 2020년도

☞ 데이터 분류 : 타자, 투수

☞ 시즌 : KBO 정규 시즌

☞ 활용변수 : 세이버 스탯 데이터

☞ 타자와 투수의 수치화 된

세이버 스탯 데이터 및 연봉 데이터

## 데이터 전처리 전

### [타자 데이터]

```
> head(ALL_Hitter_record)
```

|   | NAME | PA  | HR.  | BB.  | K.   | BB.K | IsoP  | IsoD  | BABIP | Spd | PSN  | wOBA  | wRC   | wRC/27 | wRAA | WAR. |
|---|------|-----|------|------|------|------|-------|-------|-------|-----|------|-------|-------|--------|------|------|
| 1 | 김재환  | 602 | 7.31 | 9.8  | 22.3 | 0.44 | 0.323 | 0.071 | 0.371 | 3.5 | 3.83 | 0.438 | 134.7 | 9.99   | 50.2 | 6.94 |
| 2 | 양의지  | 503 | 4.57 | 9.0  | 8.0  | 1.13 | 0.228 | 0.070 | 0.351 | 3.9 | 9.52 | 0.431 | 109.0 | 9.78   | 38.4 | 6.42 |
| 3 | 최주환  | 590 | 4.41 | 8.6  | 15.1 | 0.57 | 0.249 | 0.063 | 0.355 | 4.9 | 1.93 | 0.413 | 118.5 | 8.86   | 35.6 | 4.66 |
| 4 | 정수빈  | 112 | 1.79 | 9.8  | 11.6 | 0.85 | 0.102 | 0.061 | 0.400 | 6.2 | 2.86 | 0.402 | 21.3  | 8.57   | 5.5  | 1.06 |
| 5 | 국해성  | 28  | 0.00 | 10.7 | 17.9 | 0.60 | 0.125 | 0.095 | 0.421 | 2.7 | 0.00 | 0.402 | 5.3   | 8.46   | 1.4  | 0.15 |
| 6 | 오재일  | 477 | 5.66 | 12.6 | 25.4 | 0.50 | 0.259 | 0.094 | 0.323 | 3.8 | 1.93 | 0.384 | 82.6  | 7.36   | 15.6 | 2.85 |
| . | .    | .   | .    | .    | .    | .    | .     | .     | .     | .   | .    | .     | .     | .      | .    | .    |

## 데이터 전처리 후

### [타자 데이터]

```
> head(ALL_Data_Hitter)
```

|   | NAME | PA  | HR%  | BB%  | K%   | BB/K | IsoP  | IsoD  | BABIP | Spd | PSN  | wOBA  | wRC   | wRC/27 | wRAA | WAR+ |
|---|------|-----|------|------|------|------|-------|-------|-------|-----|------|-------|-------|--------|------|------|
| 1 | 김재환  | 602 | 7.31 | 9.8  | 22.3 | 0.44 | 0.323 | 0.071 | 0.371 | 3.5 | 3.83 | 0.438 | 134.7 | 9.99   | 50.2 | 6.94 |
| 2 | 양의지  | 503 | 4.57 | 9.0  | 8.0  | 1.13 | 0.228 | 0.070 | 0.351 | 3.9 | 9.52 | 0.431 | 109.0 | 9.78   | 38.4 | 6.42 |
| 3 | 최주환  | 590 | 4.41 | 8.6  | 15.1 | 0.57 | 0.249 | 0.063 | 0.355 | 4.9 | 1.93 | 0.413 | 118.5 | 8.86   | 35.6 | 4.66 |
| 4 | 정수빈  | 112 | 1.79 | 9.8  | 11.6 | 0.85 | 0.102 | 0.061 | 0.400 | 6.2 | 2.86 | 0.402 | 21.3  | 8.57   | 5.5  | 1.06 |
| 5 | 국해성  | 28  | 0.00 | 10.7 | 17.9 | 0.60 | 0.125 | 0.095 | 0.421 | 2.7 | 0.00 | 0.402 | 5.3   | 8.46   | 1.4  | 0.15 |
| 6 | 오재일  | 477 | 5.66 | 12.6 | 25.4 | 0.50 | 0.259 | 0.094 | 0.323 | 3.8 | 1.93 | 0.384 | 82.6  | 7.36   | 15.6 | 2.85 |
| . | .    | .   | .    | .    | .    | .    | .     | .     | .     | .   | .    | .     | .     | .      | .    | .    |

### [투수 데이터]

```
> head(ALL_Pitcher_record)
```

|   | NAME | G  | CG | SHO | GS | W  | L | SV | HLD | IP    | R  | ER | TBF | H   | X2B | X3B | HR | BB | IBB | HPB | SO  | BK | WP | ERA  | FIP  | WHIP | ERA+  | FIP+  | WAR  | WPA   | BABIP | LOB  |
|---|------|----|----|-----|----|----|---|----|-----|-------|----|----|-----|-----|-----|-----|----|----|-----|-----|-----|----|----|------|------|------|-------|-------|------|-------|-------|------|
| 1 | 린드블럼 | 26 | 0  | 0   | 26 | 15 | 4 | 0  | 0   | 168.2 | 56 | 54 | 681 | 142 | 33  | 2   | 16 | 38 | 0   | 8   | 157 | 0  | 10 | 2.88 | 4.02 | 1.07 | 175.4 | 125.2 | 6.81 | 2.06  | 0.276 | 0.80 |
| 2 | 후랭코프 | 28 | 0  | 0   | 28 | 18 | 3 | 0  | 0   | 149.1 | 64 | 62 | 621 | 118 | 17  | 1   | 12 | 55 | 1   | 22  | 134 | 0  | 15 | 3.74 | 4.61 | 1.16 | 135.3 | 110.0 | 4.12 | 1.18  | 0.270 | 0.74 |
| 3 | 이용찬  | 25 | 1  | 0   | 24 | 15 | 3 | 0  | 0   | 144.0 | 62 | 58 | 608 | 151 | 24  | 1   | 14 | 36 | 2   | 8   | 102 | 0  | 6  | 3.63 | 4.56 | 1.30 | 139.4 | 110.8 | 3.91 | 0.45  | 0.309 | 0.76 |
| 4 | 함덕주  | 62 | 0  | 0   | 0  | 6  | 3 | 27 | 3   | 67.0  | 23 | 22 | 293 | 58  | 9   | 1   | 4  | 37 | 3   | 4   | 75  | 0  | 4  | 2.96 | 4.07 | 1.42 | 171.0 | 124.8 | 2.95 | 1.58  | 0.318 | 0.81 |
| 5 | 박치국  | 67 | 0  | 0   | 0  | 1  | 5 | 3  | 17  | 67.0  | 30 | 27 | 293 | 80  | 13  | 0   | 5  | 15 | 4   | 9   | 60  | 1  | 1  | 3.63 | 3.91 | 1.42 | 139.4 | 129.6 | 1.72 | 0.27  | 0.371 | 0.76 |
| 6 | 이영하  | 40 | 0  | 0   | 17 | 10 | 3 | 0  | 2   | 122.2 | 75 | 72 | 555 | 140 | 28  | 5   | 15 | 54 | 1   | 9   | 90  | 0  | 8  | 5.28 | 5.47 | 1.58 | 95.7  | 92.1  | 1.32 | -1.25 | 0.326 | 0.70 |
| . | .    | .  | .  | .   | .  | .  | . | .  | .   | .     | .  | .  | .   | .   | .   | .   | .  | .  | .   | .   | .   | .  | .  | .    | .    | .    | .     | .     | .    | .     | .     |      |

### [투수 데이터]

```
> head(ALL_Data_Pitcher)
```

|   | NAME | G  | CG | SHO | GS | W  | L | SV | HLD | IP    | R  | ER | TBF | H   | X2B | X3B | HR | BB | IBB | HPB | SO  | BK | WP | ERA  | FIP  | ERA+ | FIP+  | WAR   | WPA  | BABIP | LOB   |      |
|---|------|----|----|-----|----|----|---|----|-----|-------|----|----|-----|-----|-----|-----|----|----|-----|-----|-----|----|----|------|------|------|-------|-------|------|-------|-------|------|
| 1 | 린드블럼 | 26 | 0  | 0   | 26 | 15 | 4 | 0  | 0   | 168.2 | 56 | 54 | 681 | 142 | 33  | 2   | 16 | 38 | 0   | 8   | 157 | 0  | 10 | 2.88 | 4.02 | 1.07 | 175.4 | 125.2 | 6.81 | 2.06  | 0.276 | 0.80 |
| 2 | 후랭코프 | 28 | 0  | 0   | 28 | 18 | 3 | 0  | 0   | 149.1 | 64 | 62 | 621 | 118 | 17  | 1   | 12 | 55 | 1   | 22  | 134 | 0  | 15 | 3.74 | 4.61 | 1.16 | 135.3 | 110.0 | 4.12 | 1.18  | 0.270 | 0.74 |
| 3 | 이용찬  | 25 | 1  | 0   | 24 | 15 | 3 | 0  | 0   | 144.0 | 62 | 58 | 608 | 151 | 24  | 1   | 14 | 36 | 2   | 8   | 102 | 0  | 6  | 3.63 | 4.56 | 1.30 | 139.4 | 110.8 | 3.91 | 0.45  | 0.309 | 0.76 |
| 4 | 함덕주  | 62 | 0  | 0   | 0  | 6  | 3 | 27 | 3   | 67.0  | 23 | 22 | 293 | 58  | 9   | 1   | 4  | 37 | 3   | 4   | 75  | 0  | 4  | 2.96 | 4.07 | 1.42 | 171.0 | 124.8 | 2.95 | 1.58  | 0.318 | 0.81 |
| 5 | 박치국  | 67 | 0  | 0   | 0  | 1  | 5 | 3  | 17  | 67.0  | 30 | 27 | 293 | 80  | 13  | 0   | 5  | 15 | 4   | 9   | 60  | 1  | 1  | 3.63 | 3.91 | 1.42 | 139.4 | 129.6 | 1.72 | 0.27  | 0.371 | 0.76 |
| 6 | 이영하  | 40 | 0  | 0   | 17 | 10 | 3 | 0  | 2   | 122.2 | 75 | 72 | 555 | 140 | 28  | 5   | 15 | 54 | 1   | 9   | 90  | 0  | 8  | 5.28 | 5.47 | 1.58 | 95.7  | 9     |      |       |       |      |

## 사용한 Python 라이브러리

```
import pandas as pd
import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt
```

[ 데이터 탐색을 위한 Python 라이브러리 ]

```
from sklearn import linear_model
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from math import sqrt
```

[ 회귀 분석을 위한 Python 라이브러리 ]

## 프로그래밍 언어 및 구현 환경

### [ 라이브러리 조건 ]

- 데이터 분석을 위해 Python 라이브러리 7가지 사용
  - ☞ 데이터 탐색 라이브러리 3가지 + 회귀 분석 라이브러리 4가지
  - ☞ 개발의 편의성을 위해 모든 라이브러리명을 요약어로 변경

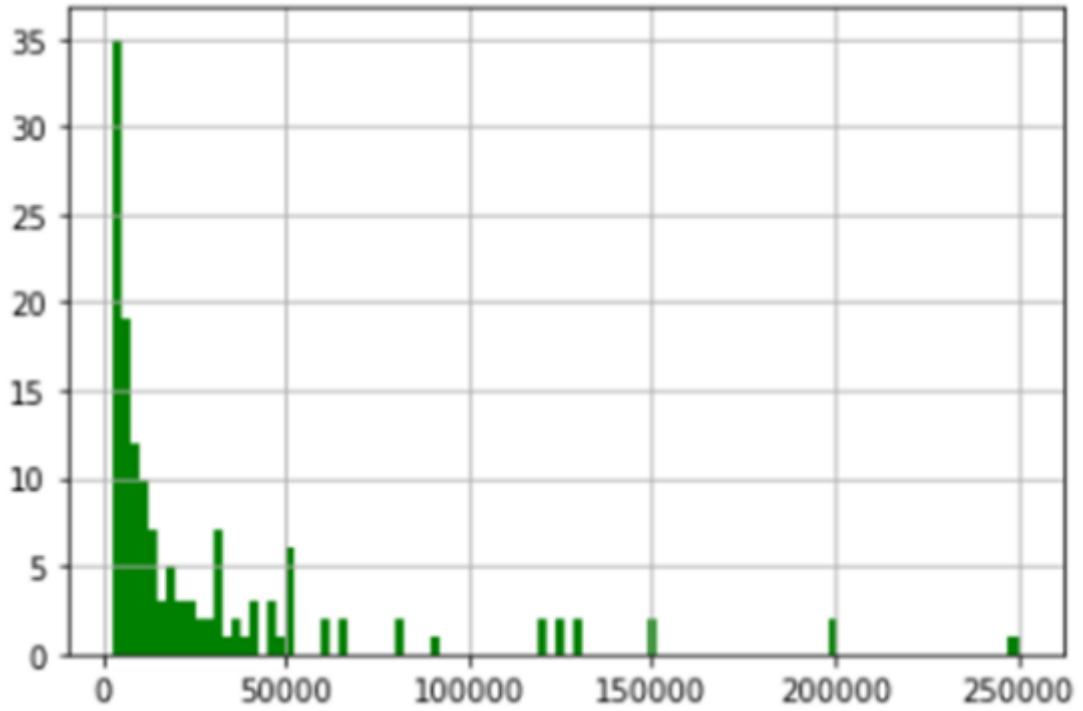
### [ CSV 파일 조건 ]

- 정제 데이터를 분석 환경으로 불러오기
  - ☞ 타자는 batter, 투수는 pitcher로 변수명 지정
- 한글 구현을 위해 CP949코드를 이용하여 Encoding
- R을 이용한 데이터 전처리 과정을 진행
  - ☞ Feature Scaling 외 추가 데이터 처리 필요성 없음

## [ 타자의 종속 변수 데이터 정보 ]

|          |        |               |
|----------|--------|---------------|
| 전체 데이터   | count  | 143.000000    |
| 평균 값     | mean   | 28878.321678  |
| 표준편차     | std    | 42154.065414  |
| 최소값      | min    | 2700.000000   |
| 제 1 사분위수 | 25%    | 5500.000000   |
| 제 2 사분위수 | 50%    | 12000.000000  |
| 제 3 사분위수 | 75%    | 32000.000000  |
| 최대값      | max    | 250000.000000 |
| 변수명      | Name:  | SALARY(2020)  |
| 데이터 타입   | dtype: | float64       |

## [ 타자의 종속 변수 데이터 그래프 ]



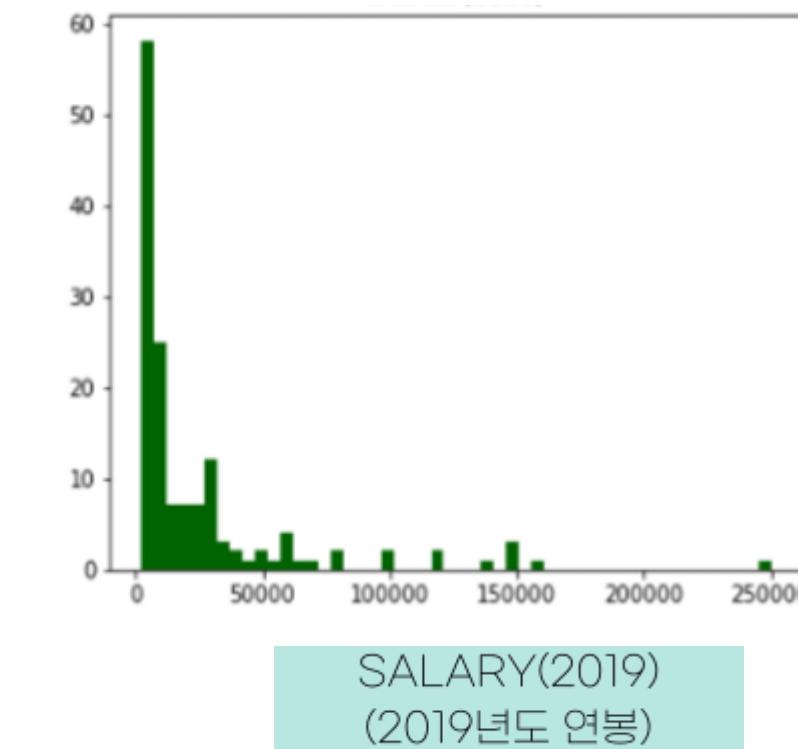
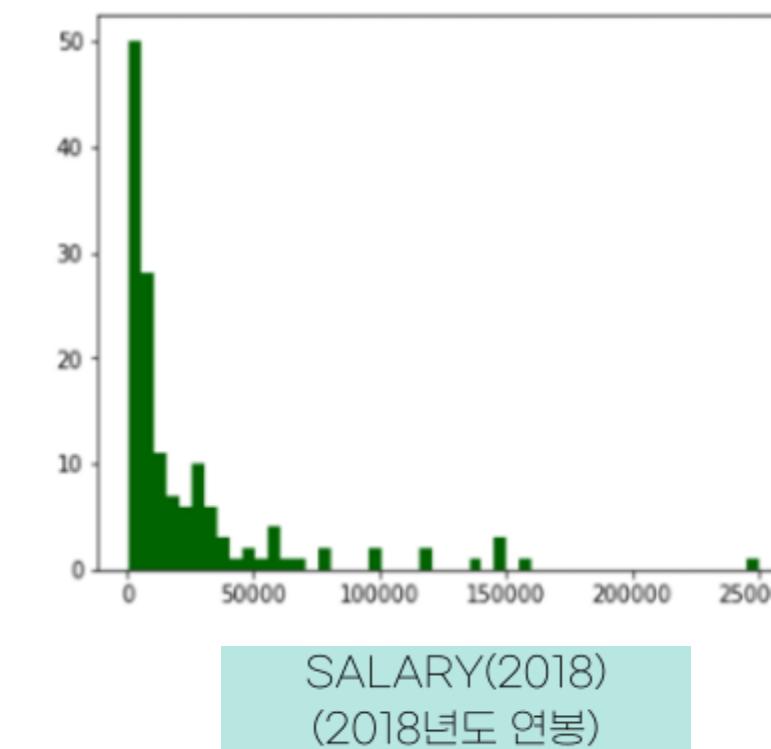
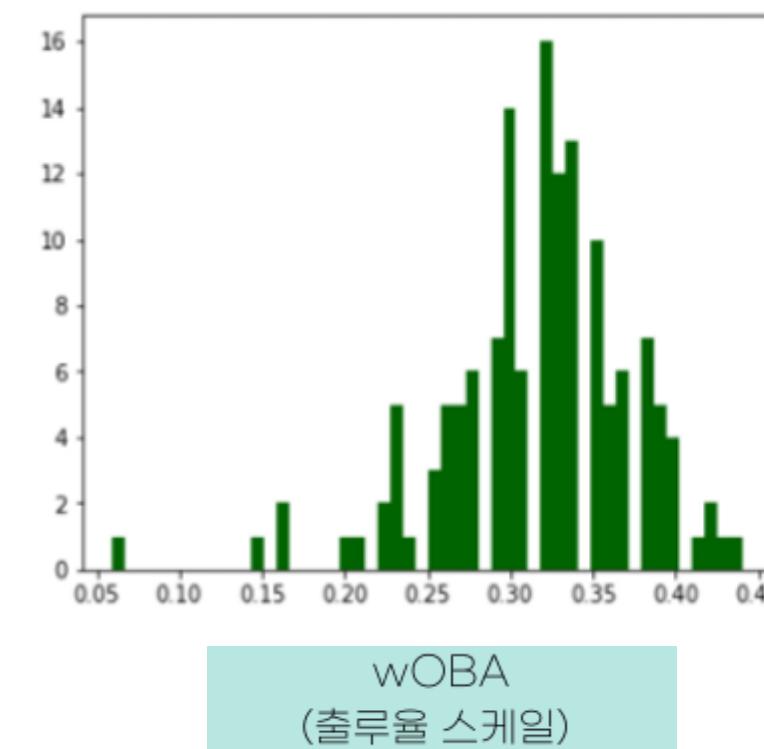
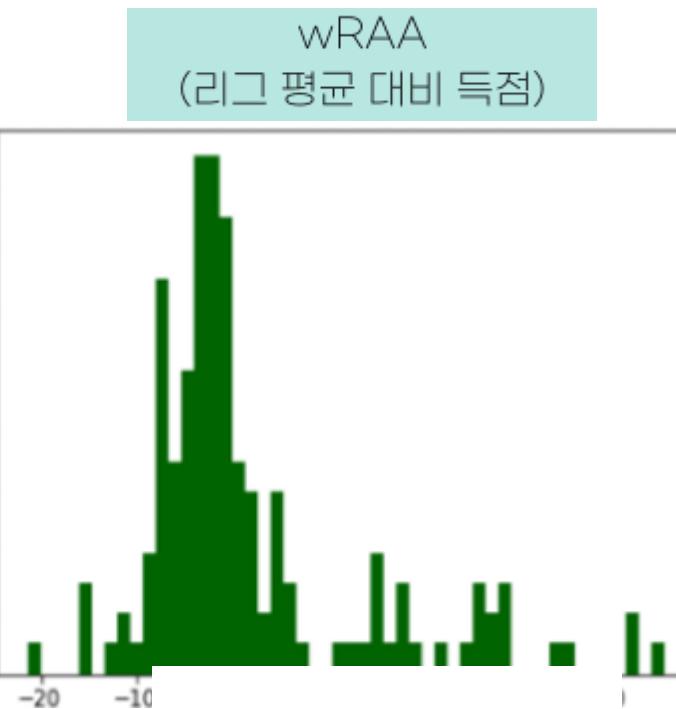
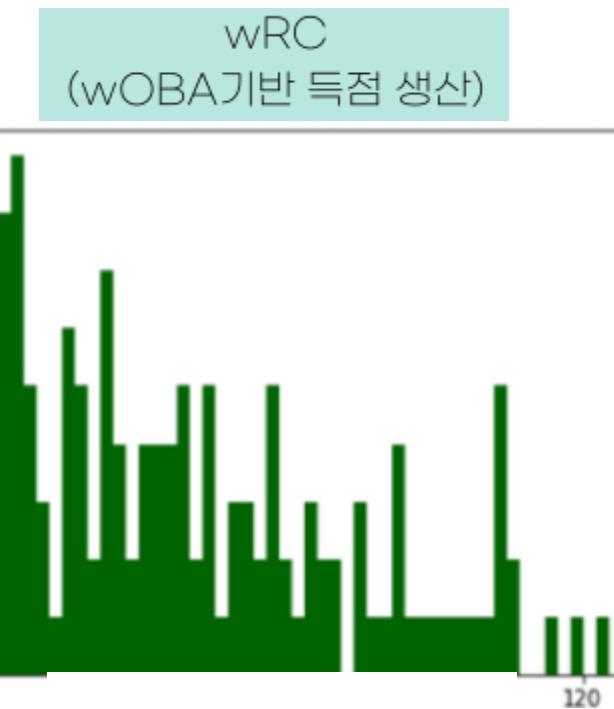
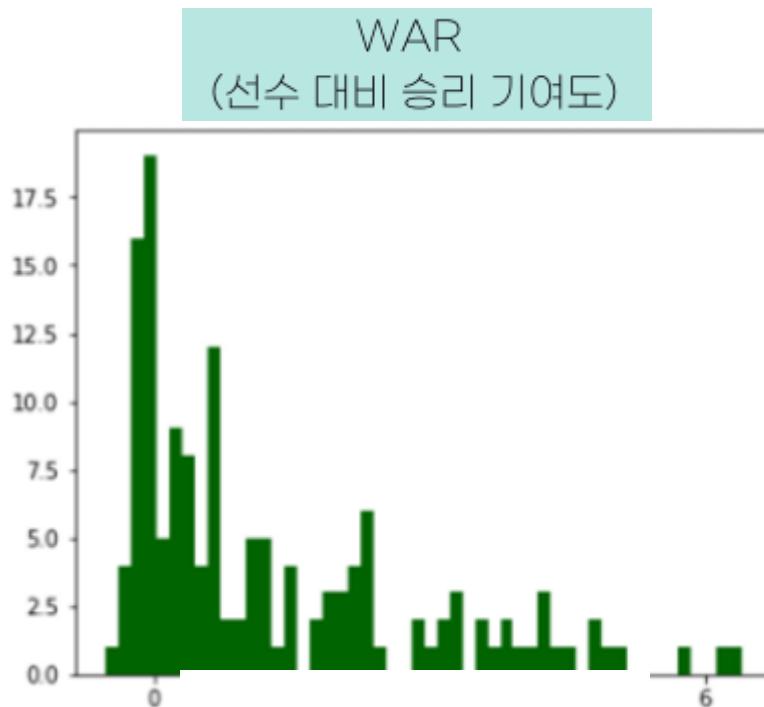
## 종속 변수 데이터

- 종속 변수 데이터 정보
  - ☞ 예측 대상에 대한 정보 확인
- 예측 대상의 종속 변수 데이터 분포 파악
  - ☞ 실제 2020년도 연봉 데이터 확인(변수명:SALARY(2020))
  - ☞ 그래프를 통해 파악을 용이하게 함

\*실제 데이터와 앞으로의 분석 예측을 비교하는 지표로 사용

## 종속 변수 데이터

[ 타자의 독립 변수 데이터 그래프 ]

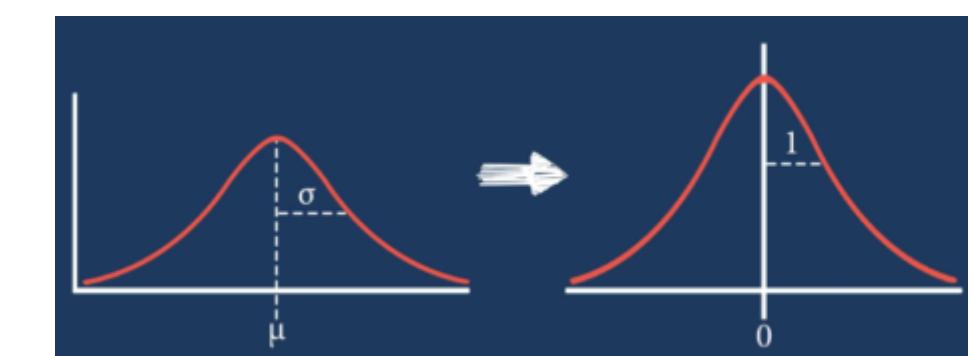


- 회귀분석에 사용할 6개 독립변수 확인
  - ☞ 예측 대상에 대한 정보 확인
- 각 독립 변수마다의 편차 차이 극복 필요
  - ☞ Feature Scaling



큰 값을 지닌 독립 변수가 종속 변수에 과대한 영향

- 독립 변수의 범위를 정규화시키는 것
- Z-score 표준화 :  $\frac{((x) - (x\text{의 평균}))}{x\text{의 표준편차}}$



[Feature Scaling 이후 타자의 데이터 요약표]

| NAME  | WAR       | wRC       | wRAA      | wOBA      | SALARY(2018) | SALARY(2019) | y      |
|-------|-----------|-----------|-----------|-----------|--------------|--------------|--------|
| 0 강경학 | -0.376215 | -0.401207 | -0.355267 | 0.056818  | -0.503903    | -0.505972    | 7800   |
| 1 강민호 | 0.596649  | 0.480522  | -0.115339 | 0.402564  | 1.950849     | 1.950861     | 125000 |
| 2 강백호 | 1.575711  | 1.859950  | 2.188631  | 1.439800  | -0.584686    | -0.586823    | 21000  |
| 3 강진성 | -0.438181 | -0.413495 | -0.134237 | -0.116054 | -0.576868    | -0.578999    | 3800   |
| 4 고종욱 | -0.394805 | 0.170227  | -0.451403 | 0.056818  | -0.368397    | -0.370351    | 17000  |

## Feature Scaling

- Z-score 표준화를 통해 Feature Scaling
- ☞ 각 변수의 단위를 상대적 값을 표현할 수 있는 수치로 변경

[타자의 회귀계수]

|              |                 |              |                 |
|--------------|-----------------|--------------|-----------------|
| WAR          | -10450.95388302 | wRC          | 47773.55990713  |
| wRAA         | 18671.74691477  | wOBA         | 1809.21296527   |
| SALARY(2018) | 1583.25723169   | SALARY(2019) | -12090.15447108 |

## 회귀계수 도출

- 데이터셋을 학습 데이터와 검증 데이터로 분할
  - ☞ 학습 데이터를 통해 회귀분석 모델 구축을 위한 학습 데이터셋으로 사용
  - ☞ 검증데이터를 통해 회귀분석 모델의 학습 정확도를 평가하기 위한 데이터셋으로 사용
- 모델 학습을 위한 회귀 계수 도출
  - ☞ 데이터수의 한계를 반영하여 학습 데이터를 0.5로 설정

## 예측 모델의 변수 평가

- 변수의 유의미를 파악하기 위해 예측 모델의 변수 평가를 진행
- 결정 계수는 모두 1에 근접할수록 분석 예측도가 높다고 간주 가능
  - ☞ 결정계수 : 분석의 예측도를 평가하는 지표
  - 표를 기반으로 결정계수는 0.900, 수정결정계수는 0.890
- 분석의 예측도가 높다고 간주할 수 있음.
- WAR의  $P > |t|$  값이 0.023 ☞ 해당 변수가 가장 유의미한 핵심 변수
  - ☞  $|t|$  수치는 각 변수의 검정 통계량이 얼마나 유의미한지 의미
  - ☞ p-value ( $= P > |t|$ )의 값이 0.05 이하면 모델의 예측이 유의미

[타자의 검증데이터표]

### OLS Regression Results

| Dep. Variable:    | y                | R-squared:          | 0.900     |       |           |          |
|-------------------|------------------|---------------------|-----------|-------|-----------|----------|
| Model:            | OLS              | Adj. R-squared:     | 0.890     |       |           |          |
| Method:           | Least Squares    | F-statistic:        | 95.74     |       |           |          |
| Date:             | Tue, 01 Jun 2021 | Prob (F-statistic): | 4.94e-30  |       |           |          |
| Time:             | 16:26:38         | Log-Likelihood:     | -787.16   |       |           |          |
| No. Observations: | 71               | AIC:                | 1588.     |       |           |          |
| Df Residuals:     | 64               | BIC:                | 1604.     |       |           |          |
| Df Model:         | 6                |                     |           |       |           |          |
| Covariance Type:  | nonrobust        |                     |           |       |           |          |
|                   | coef             | std err             | t         | P> t  | [0.025    | 0.975]   |
| const             | 3.016e+04        | 1991.822            | 15.144    | 0.000 | 2.62e+04  | 3.41e+04 |
| SALARY(2018)      | -1.045e+04       | 7.24e+04            | -0.144    | 0.886 | -1.55e+05 | 1.34e+05 |
| SALARY(2019)      | 4.777e+04        | 7.24e+04            | 0.660     | 0.512 | -9.69e+04 | 1.92e+05 |
| WAR               | 1.867e+04        | 8036.644            | 2.323     | 0.023 | 2616.705  | 3.47e+04 |
| wOBA              | 1809.2130        | 3976.419            | 0.455     | 0.651 | -6134.597 | 9753.023 |
| wRAA              | 1583.2572        | 4428.036            | 0.358     | 0.722 | -7262.761 | 1.04e+04 |
| wRC               | -1.209e+04       | 7969.576            | -1.517    | 0.134 | -2.8e+04  | 3830.904 |
| Omnibus:          | 64.511           | Durbin-Watson:      | 1.915     |       |           |          |
| Prob(Omnibus):    | 0.000            | Jarque-Bera (JB):   | 754.552   |       |           |          |
| Skew:             | 2.324            | Prob(JB):           | 1.42e-164 |       |           |          |
| Kurtosis:         | 18.279           | Cond. No.           | 120.      |       |           |          |

## 예측 모델 평가

- ✓ ① 수정 결정 계수 측정 방법 ② 평균 제곱근의 편차 도출 방법
- ☞ 그러나 평균 제곱근 편차에 대한 기준이 명확하지 않음
- ☞ 학습 데이터와 검증 데이터의 수정 결정 계수를 측정하는  
수정 결정 계수 측정 방법 사용

- 두 데이터 간의 수정 결정 계수의 값이 차이가 작을수록  
예측 모델의 정확도가 높다고 판단
- ☞ 학습 데이터의 수정 결정계수는 0.89975, 검증 데이터의 수정 결정계수는 0.64027
- ☞ 모형의 예측도가 정확하다고 추론

### R2 score (결정계수)

$$R^2 = 1 - \frac{\sum_{i=1}^n (y^i - \hat{y}^i)^2}{\sum_{i=1}^n (y^i - \bar{y})^2}$$

- 분산을 기반으로 성능을 평가함
- 상대적으로 성능이 어느 정도인지 판단할 수 있음.
- 1에 가까워 질수록 성능이 좋은 모델임.
- Test set 과 Train set 간의 값의 차이가 작을수록  
예측 모델의 정확도가 높음

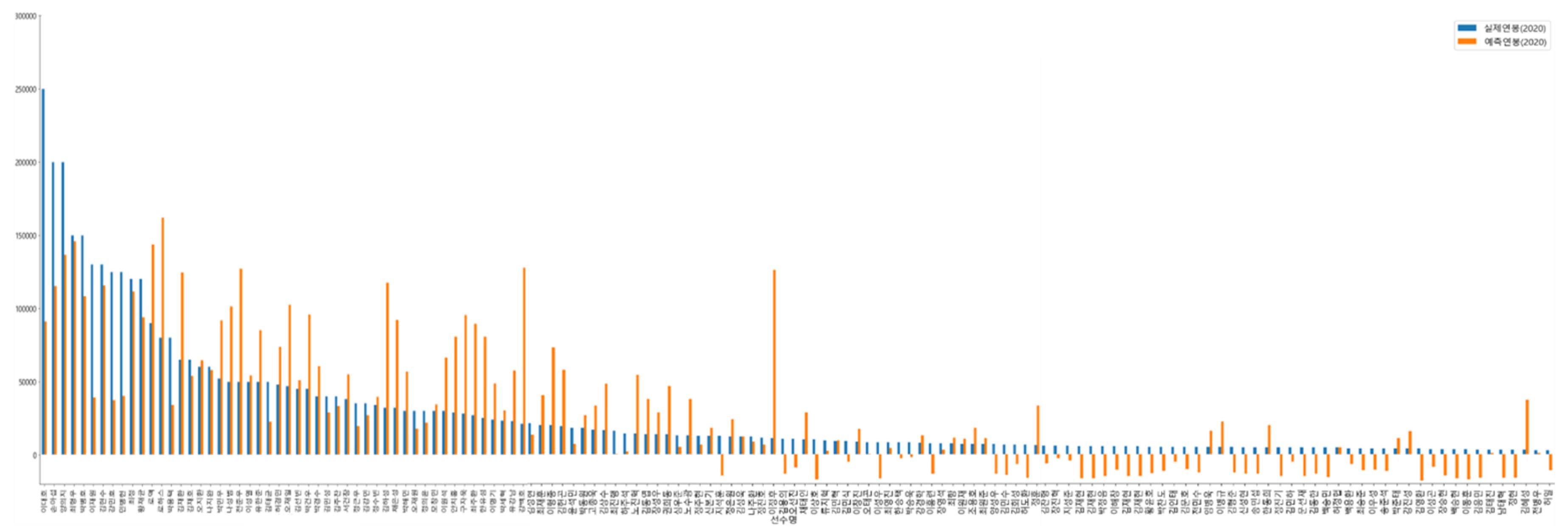
### RMSE score (평균 제곱근 편차)

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2}$$

- MSE(실제값에서 예측값을 뺀 제곱의 평균)에 루트를 씌운 값
- 제곱한 값에 다시 루트를 쓰우기 때문에
- 오차의 왜곡이라는 단점 보완 가능

## 타자 데이터 분석 결과 그래프

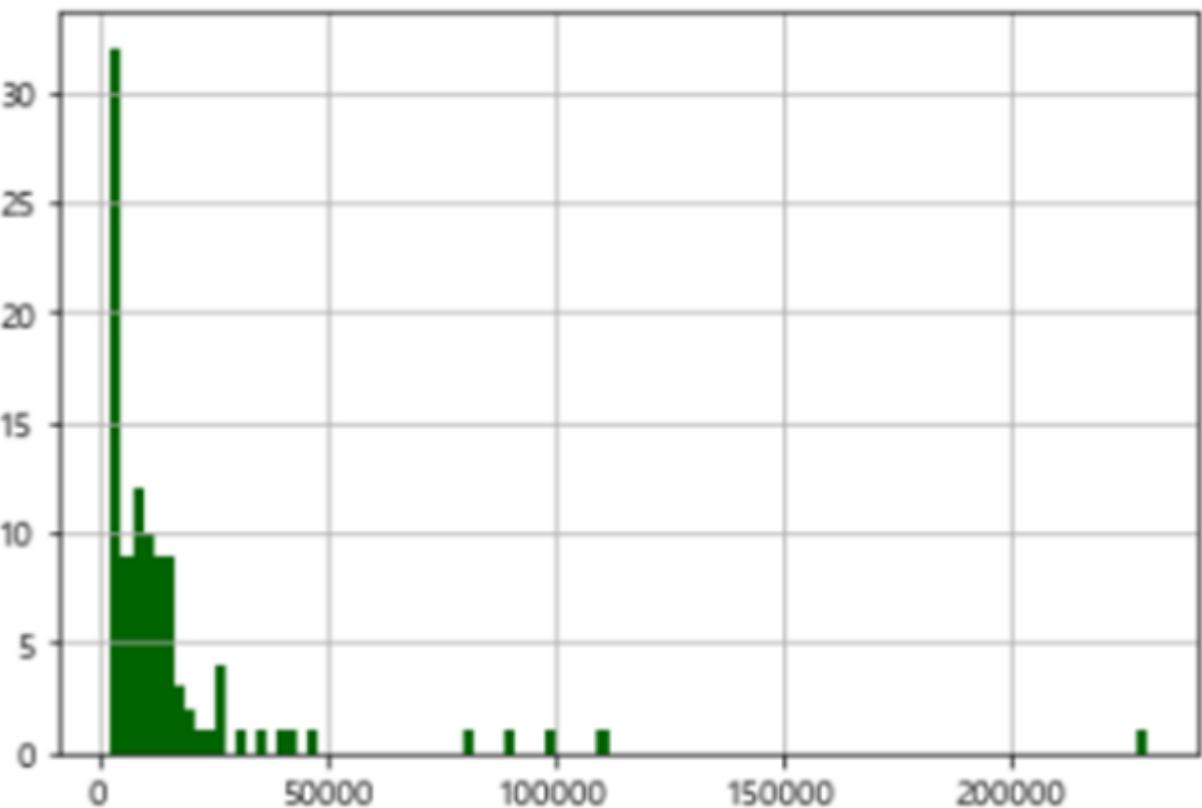
상황분석 | 문제도출 | 데이터분석 | 기대효과  
3) 데이터 분석 - 타자



## [ 투수의 종속 변수 데이터 정보 ]

```
전체 데이터      count      102.000000
평균 값          mean      16382.352941
표준편차        std       28439.546486
최소값          min       2700.000000
제 1 사분위수    25%      4200.000000
제 2 사분위수    50%      8900.000000
제 3 사분위수    75%      15000.000000
최대값          max      230000.000000
변수명          Name: SALARY(2020)
데이터 타입      dtype: float64
```

## [ 투수의 종속 변수 데이터 그래프 ]

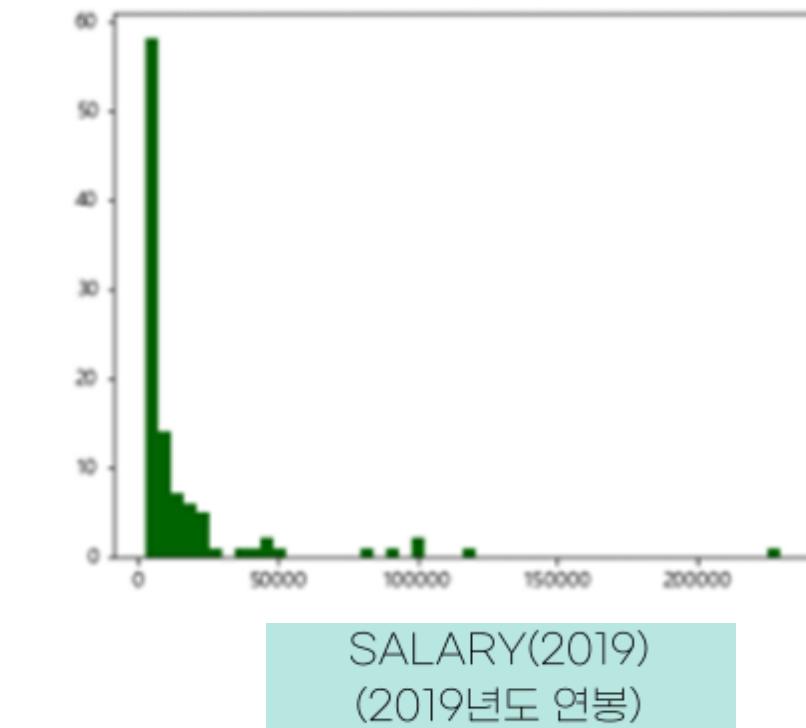
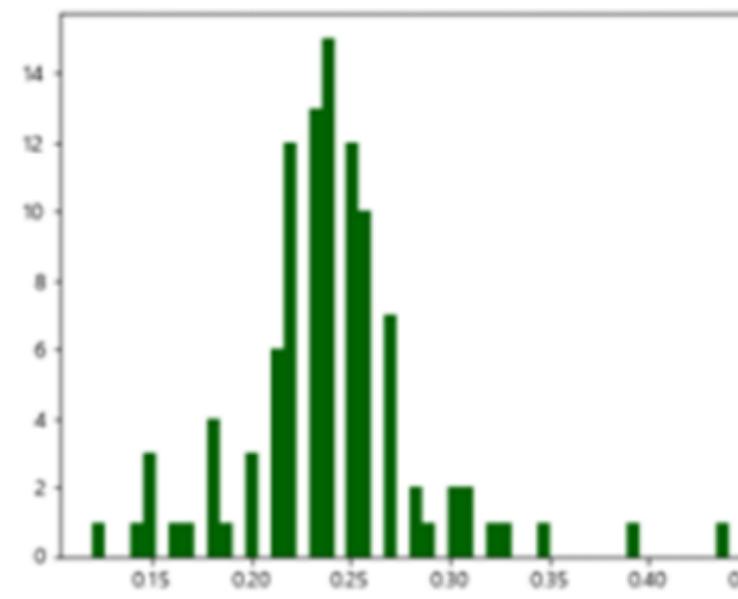
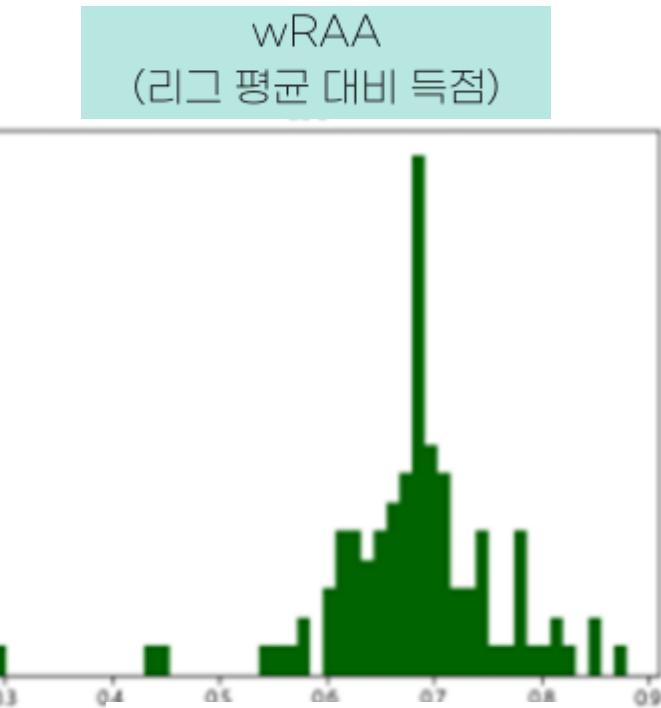
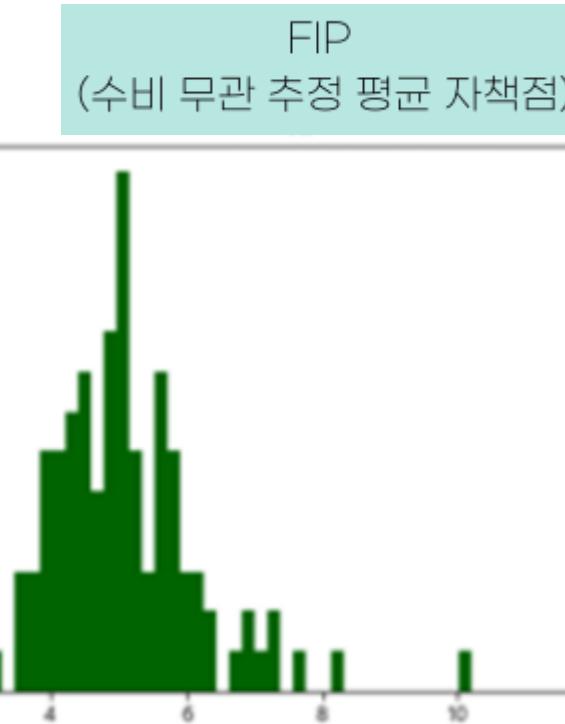
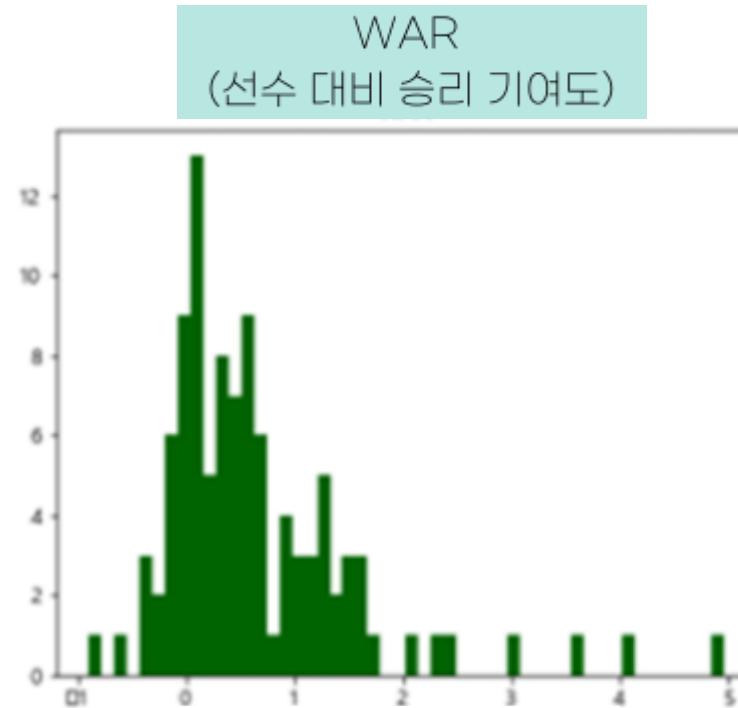


## 종속 변수 데이터

- 종속 변수 데이터 정보
  - ☞ 예측 대상에 대한 정보 확인
- 예측 대상의 종속 변수 데이터 분포 파악
  - ☞ 실제 2020년도 연봉 데이터 확인(변수명:SALARY(2020))
  - ☞ 그래프를 통해 파악을 용이하게 함

\*실제 데이터와 앞으로의 분석 예측을 비교하는 지표로 사용

## [ 투수의 독립 변수 데이터 그래프 ]



## 종속 변수 데이터

- 회귀분석에 사용할 6개 독립변수 확인
  - 예측 대상에 대한 정보 확인
- 각 독립 변수마다의 편차 차이 극복 필요
  - Feature Scaling

[Feature Scaling 이후 투수의 데이터 요약표]

|   | NAME | WAR       | FIP       | LOB       | BABIP     | SALARY(2018) | SALARY(2019) | y     |
|---|------|-----------|-----------|-----------|-----------|--------------|--------------|-------|
| 0 | 강동연  | -0.683585 | 0.874631  | 0.469024  | -0.421414 | -0.433687    | -0.434398    | 3400  |
| 1 | 강윤구  | -0.651038 | -0.175467 | -0.257666 | 0.442591  | -0.246190    | -0.247566    | 15500 |
| 2 | 고우석  | 0.911198  | -0.735008 | 0.469024  | -0.637416 | -0.394214    | -0.395065    | 22000 |
| 3 | 구승민  | 0.466395  | -0.428410 | 0.590139  | -0.421414 | -0.413950    | -0.414732    | 8000  |
| 4 | 구창모  | 2.614471  | -0.865312 | 1.437944  | -0.637416 | -0.338294    | -0.339343    | 18000 |

## Feature Scaling

- z-score 표준화를 통해 Feature Scaling
  - ☞ 각 변수의 단위를 상대적 값을 표현할 수 있는 수치로 변경

## 회귀계수 도출

[투수의 회귀계수]

|                     |                |                     |                |
|---------------------|----------------|---------------------|----------------|
| <b>WAR</b>          | 128.50705397   | <b>FIP</b>          | 906.17560653   |
| <b>LOB%</b>         | -486.0567842   | <b>BABIP</b>        | -2257.07667265 |
| <b>SALARY(2018)</b> | 28436.78053828 | <b>SALARY(2019)</b> | 7430.19818042  |

- 데이터셋을 학습 데이터와 검증 데이터로 분할
  - ☞ 학습 데이터를 통해 회귀분석 모델 구축을 위한 학습 데이터셋으로 사용
  - ☞ 검증데이터를 통해 회귀분석 모델의 학습 정확도를 평가하기 위한 데이터셋으로 사용
- 모델 학습을 위한 회귀 계수 도출
  - ☞ 데이터수의 한계를 반영하여 학습 데이터를 0.5로 설정

[투수의 검증데이터표]

## 예측 모델의 변수 평가

- 변수의 유의미를 파악하기 위해 예측 모델의 변수 평가를 진행
- 결정 계수는 모두 1에 근접할수록 분석 예측도가 높다고 간주 가능
  - ☞ 결정계수 : 분석의 예측도를 평가하는 지표  
표를 기반으로 결정계수는 0.905, 수정결정계수는 0.892
- WAR의 P > |t| 값이 0.000 ☞ 해당 변수가 가장 유의미한 핵심 변수
  - ☞ |t| 수치는 각 변수의 검정 통계량이 얼마나 유의미한지 의미

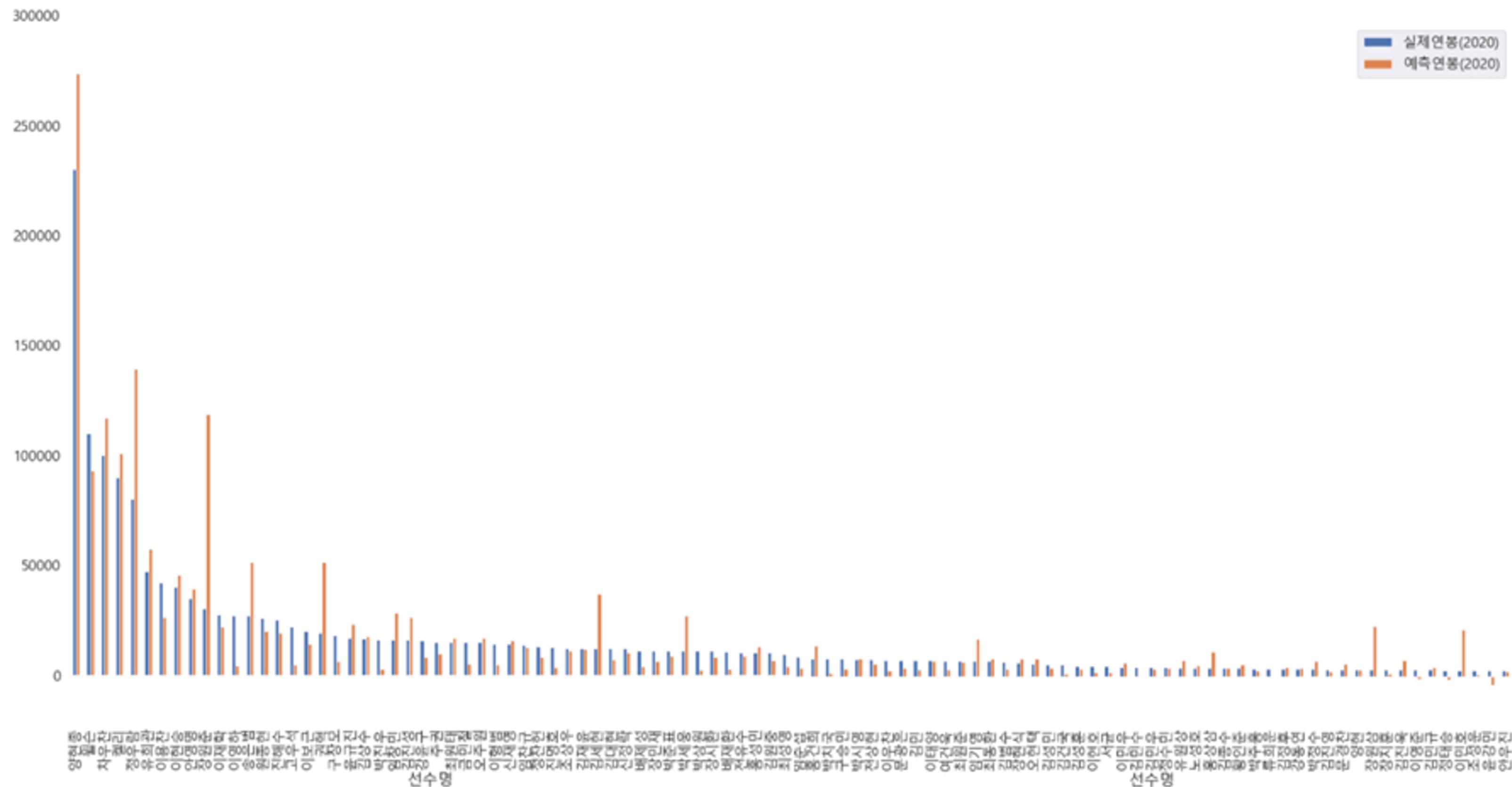
## 예측 모델 평가

- 학습 데이터의 수정 결정계수는 0.89975, 검증 데이터의 수정 결정계수는 0.64027
  - ☞ 모형의 예측도가 정확하다고 추론

| Dep. Variable:    | y                | R-squared:          | 0.905    |       |           |          |
|-------------------|------------------|---------------------|----------|-------|-----------|----------|
| Model:            | OLS              | Adj. R-squared:     | 0.892    |       |           |          |
| Method:           | Least Squares    | F-statistic:        | 69.84    |       |           |          |
| Date:             | Thu, 03 Jun 2021 | Prob (F-statistic): | 7.42e-21 |       |           |          |
| Time:             | 11:47:14         | Log-Likelihood:     | -508.13  |       |           |          |
| No. Observations: | 51               | AIC:                | 1030.    |       |           |          |
| Df Residuals:     | 44               | BIC:                | 1044.    |       |           |          |
| Df Model:         | 6                |                     |          |       |           |          |
| Covariance Type:  | nonrobust        |                     |          |       |           |          |
|                   | coef             | std err             | t        | P> t  | [0.025    | 0.975]   |
| const             | 1.797e+04        | 828.903             | 21.676   | 0.000 | 1.63e+04  | 1.96e+04 |
| BABIP             | 128.5071         | 1021.817            | 0.126    | 0.900 | -1930.829 | 2187.843 |
| FIP               | 906.1756         | 1059.712            | 0.855    | 0.397 | -1229.534 | 3041.885 |
| LOB               | -486.0568        | 1044.093            | -0.466   | 0.644 | -2590.288 | 1618.174 |
| SALARY(2018)      | -2257.0767       | 8.7e+04             | -0.026   | 0.979 | -1.78e+05 | 1.73e+05 |
| SALARY(2019)      | 2.844e+04        | 8.73e+04            | 0.326    | 0.746 | -1.48e+05 | 2.04e+05 |
| WAR               | 7430.1982        | 1264.034            | 5.878    | 0.000 | 4882.705  | 9977.691 |
| Omnibus:          | 14.873           | Durbin-Watson:      | 2.076    |       |           |          |
| Prob(Omnibus):    | 0.001            | Jarque-Bera (JB):   | 23.594   |       |           |          |
| Skew:             | -0.887           | Prob(JB):           | 7.53e-06 |       |           |          |
| Kurtosis:         | 5.821            | Cond. No.           | 265.     |       |           |          |

# 투수 데이터 분석 결과 그래프

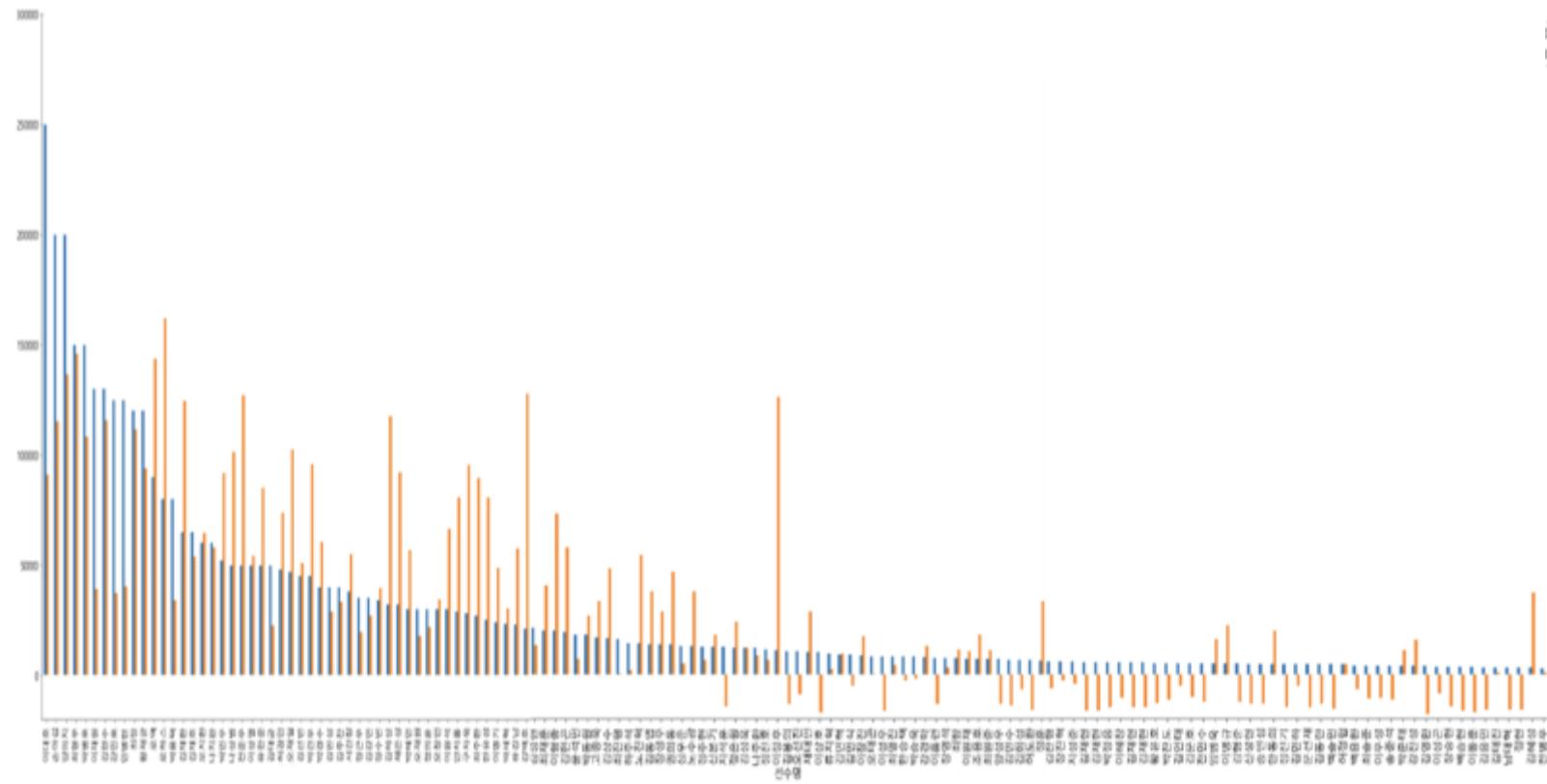
상황분석 | 문제도출 | 데이터분석 | 기대효과  
3) 데이터 분석 - 투수



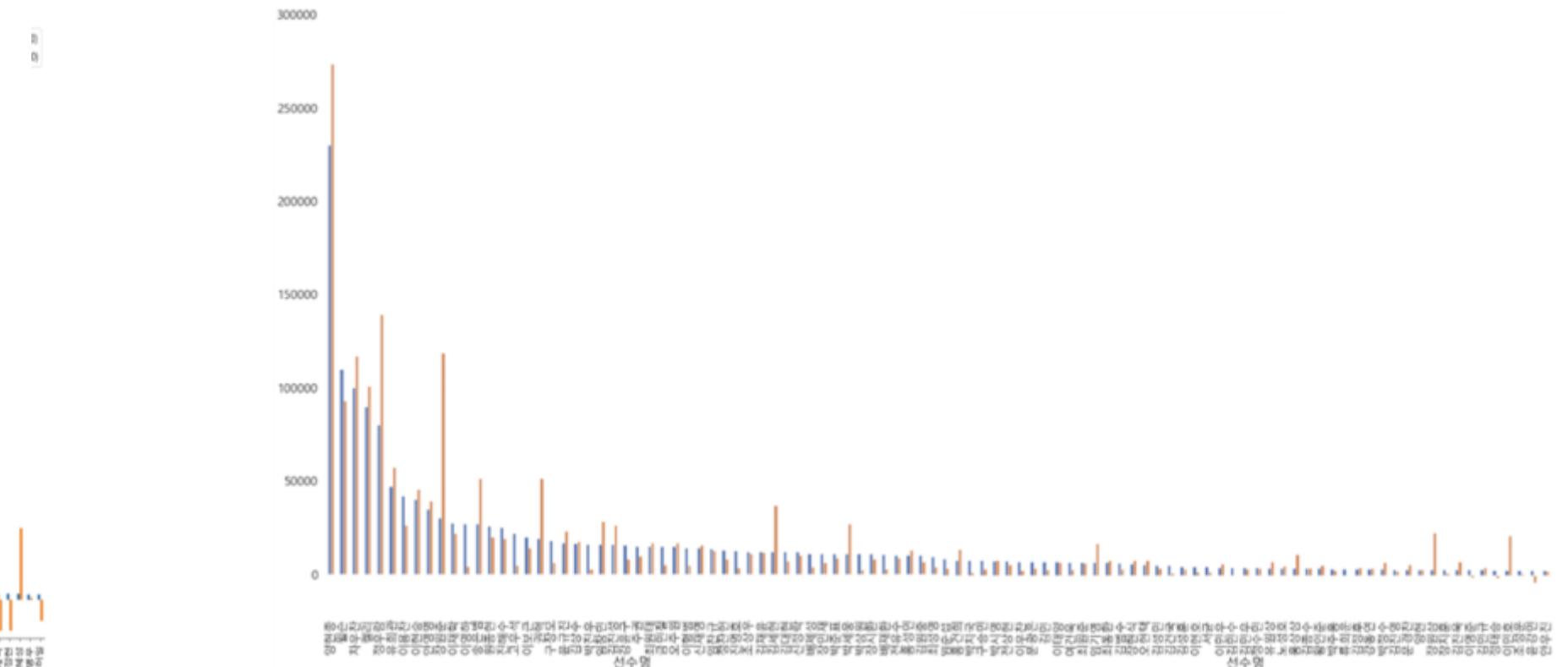
## 타자 연봉 예측 모델보다 높은 정확도를 가지는 투수 연봉 예측 모델

- 예측 연봉 (2020년)
- 실제 연봉 (2020년)

[타자 연봉 예측 모델]



[투수 연봉 예측 모델]



- 타자 예측 모델의  $WAR P > |t|$  의 값 : 0.023

- 투수 예측 모델의  $WAR P > |t|$  의 값 : 0.00

☞  $P > |t|$  의 값이 낮을수록 독립 변수가 영향력 ○

☞ 투수 연봉 예측 모델이 타자 연봉 예측 모델보다 데이터 수는 적지만 투수 연봉 예측 모델이 정확도가 높다.

## 전관예우 등의 현실적인 한계로 인해 연봉 예측 모델의 값이 음수(이상치)로 나오는 경우가 존재

- 3개년 간의 선수 개인 세이버 스탯과 기존 2개년의 연봉 데이터라는 객관적인 수치만을 가지고 모델 설계
- ☞ 그러나 전관예우, 선수에 대한 기대감 등 주관적인 평가가 포함될 수 있는 '인사'인 연봉 예측과는 다소 차이가 발생 가능

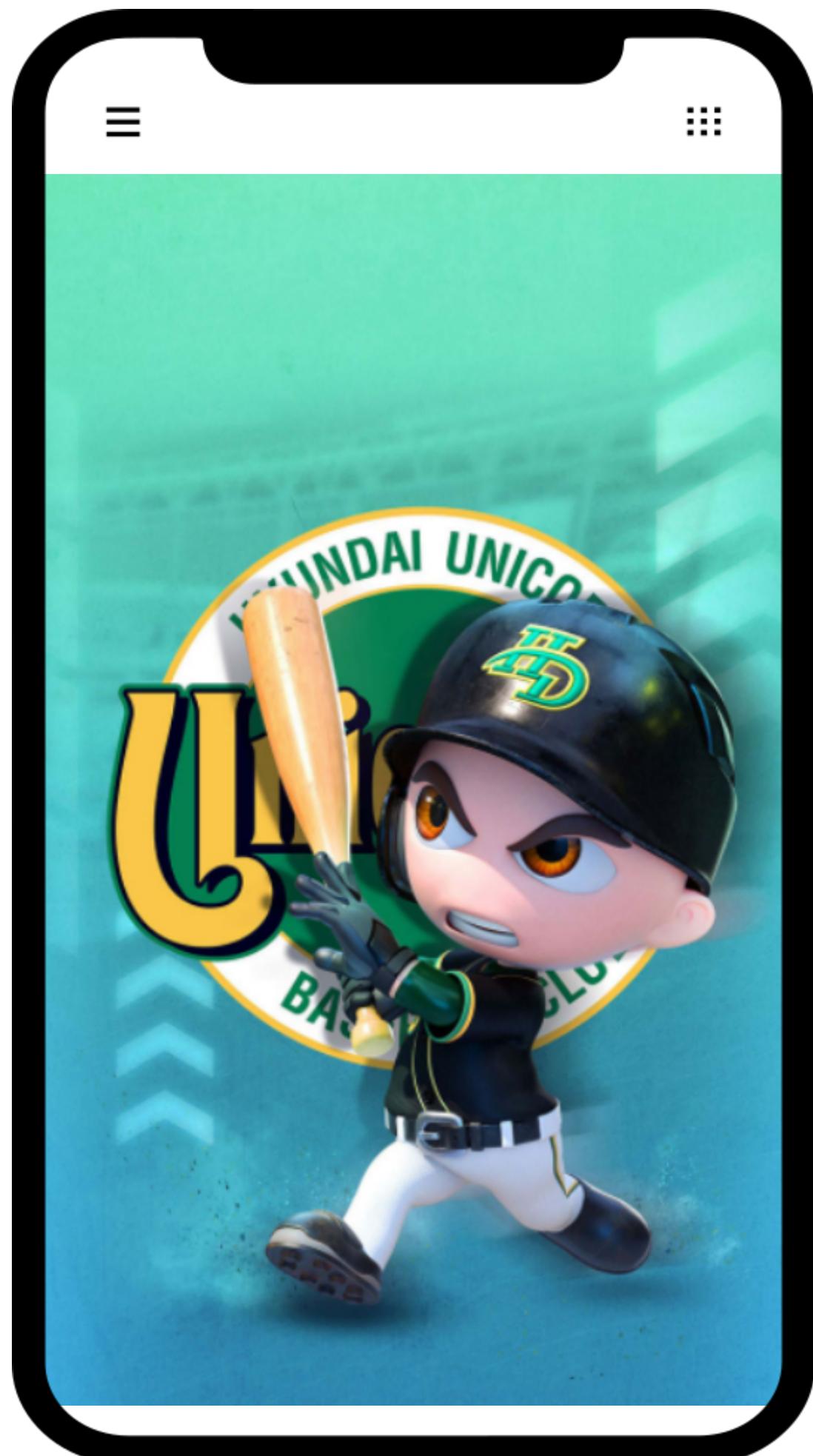
☞ 연봉 협상에서의 합의 기준점으로 사용 가능





## 연봉 협상에서의 기준점 제시합니다.

- 최근에도 구단과 선수 간 연봉 협상이 결렬되어 한국프로야구위원회(KBO)에 연봉 조정 신청을 한 경우가 발생
  - ☞ 그러나 최근 사례가 19년만에 최초로 선수가 승리할만큼 선수 측의 승리는 어려웠고, 연봉 '조정'이 아닌 연봉 '결정'이라는 대중의 비판 존재
  - ☞ 객관적인 수치를 기반으로 한 연봉 결정 모델을 사용함으로써 구단과 선수 모두 납득할 수 있는 기준점 제시 가능
  - ☞ 행정적, 절차적 비용의 낭비를 줄이고, 선수와 구단의 편의 도모 가능



## 모델 알고리즘을 통해 E-Sports 산업 발전을 도모할 수 있습니다.



'컴투스 프로야구', '마구마구' 등 게임 개발자에게  
선수의 개인의 데이터 및 알고리즘을 제공

☞ 개발 시간 단축 및 개발 비용 절약



PC 이용자들에게 현실성이 가미된  
흥미 요소를 제공



KBO 리그 10개 구단 팀별 팬들에게 즐거움 선사

## 제한점

- 포수, 경기장, 당일의 날씨와 같은 외부 변수들이 선수들의 경기력에 영향을 미칠 가능성이 큼
- 현재 사용한 변수 중 WAR을 제외한 세이버 스탯변수가 연봉을 결정하는 완벽한 변수로 간주하기에는 한계

[ 날씨 변수가 경기에 영향을 미치는 그래프 ]

