주제 : 최동원 선수가 현역 선수라면 연봉이 얼마일까?

목차

과제 소개

- 1. 최동원 선수의 데이터 및 2015~2020년까지의 전체 선수 데이터 수집

- 2. 2015~2020년 선수들과 연봉이 결정되는 해인 2016~2021년 연봉 데이터 수집 및 기존 데이터와 매핑

- 3. 최동원 선수 연봉 예측(시즌별)

과제 소개

- 1. 최동원 선수의 데이터 및 2015~2020년까지의 전체 선수 데이터 수집

- 2. 2015~2020년 선수들과 연봉이 결정되는 해인 2016~2021년 연봉 데이터 수집 및 기존 데이터와 매핑

- 데이터 전처리

- 3. 최동원 선수 연봉 예측(시즌별)

- 데이터 분석 : 영향력 있는 결정 변수만 고려하여 예측

들어가기 전 설정

1. 다음 해의 연봉은 올해의 활약에 따라서 결정된다.

- -> 포스트 시즌을 제외한 정규 시즌 데이터만으로 추론
- -> 꾸준함을 반영하기는 어렵다는 단점이 있음
- -> 복잡한 KBO 규약을 적용해서 학습하는 것은 시간이 더 오래 걸릴 일
- ->(연봉조정신청 및 FA 제도를 고민하기에는, 이전 시기와 현재의 체로를 변수화해야 하기 때문에 일단은 배제)

2. 선발/마무리 투수의 구분은 그 해의 경기 수 대비 선발 출장 경기 수를 비교하여 결정한다.

- -> 규정이닝 선수 표본수의 문제

- -> ‘스윙맨 ' 보직의 존재

최동원 선수 : 어떤 선수인가?

- 현역 시절 선동열과 함께 한국프로야구를 대표한 양대산맥으로 손꼽히는 투수이자 롯데 자이언츠를 상징하는 선수

- 아마 시절에 이미 상상을 초월하는 혹사에 시달린 후 프로에 데뷔했음에도 프로에서 뚜렷한 족적을 남겼다

연도	소속팀	경기	이닝	승	패	세	승률	ERA	피안타	피홈런	4사구	탈삼진	실점	자책점	WHIP
1983	롯데	**38** (5위)	**208⅔** (5위)	9	16	4	0.360	2.89	202	17	59	**148** (4위)	89	67	1.21
1984		**51** (2위)	**284⅔** (1위)	**27** (1위)	13	**6** (5위)	**0.675** (4위)	**2.40** (4위)	229	18	82	**223** (1위)[36] [37]	91	76	1.04
1985		**42** (5위)	**225** (4위)	**20** (3위)	9	**8** (3위)	**0.690** (3위)	**1.92** (2위)	170	7	49	**161** (2위)	60	48	0.94
1986		39	**267** (1위)	**19** (2위)	14	2	0.576	**1.55** (2위)	204	7	61	**208** (2위)	60	46	0.97
1987		32	**224** (2위)	**14** (4위)	12	2	0.538	2.81	218	6	68	**163** (1위)	80	70	1.25
1988		16	83⅓	7	3	3	0.700	2.05	77	4	25	**83** (4위)	24	19	1.21
연도	소속팀	경기	이닝	승	패	세	승률	ERA	피안타	피홈런	4사구	탈삼진	실점	자책점	WHIP
1989	삼성	8	30	1	2	0	0.333	2.10	36	2	19	9	12	7	1.80
1990		22	92	6	5	1	0.545	5.28	113	9	56	24	62	54	1.82
KBO 리그 통산 (8시즌)		248	1414⅔	103	74	26	0.582	**2.46** (2위) [38]	1249	70	419	1019	478	387	1.15

출처 : 스탯티즈 (http://www.statiz.co.kr/main.php)

연도	소속팀	경기	이닝	승	패	세	승률	ERA	피안타	피홈런	4사구	탈삼진	실점	자책점	WHIP
1983	롯데	**38** (5위)	**208⅔** (5위)	9	16	4	0.360	2.89	202	17	59	**148** (4위)	89	67	1.21
1984		**51** (2위)	**284⅔** (1위)	**27** (1위)	13	**6** (5위)	**0.675** (4위)	**2.40** (4위)	229	18	82	**223** (1위)[36] [37]	91	76	1.04
1985		**42** (5위)	**225** (4위)	**20** (3위)	9	**8** (3위)	**0.690** (3위)	**1.92** (2위)	170	7	49	**161** (2위)	60	48	0.94
1986		39	**267** (1위)	**19** (2위)	14	2	0.576	**1.55** (2위)	204	7	61	**208** (2위)	60	46	0.97
1987		32	**224** (2위)	**14** (4위)	12	2	0.538	2.81	218	6	68	**163** (1위)	80	70	1.25
1988		16	83⅓	7	3	3	0.700	2.05	77	4	25	**85** (4위)	24	19	1.21
연도	소속팀	경기	이닝	승	패	세	승률	ERA	피안타	피홈런	4사구	탈삼진	실점	자책점	WHIP
1989	삼성	8	30	1	2	0	0.333	2.10	36	2	19	9	12	7	1.80
1990		22	92	6	5	1	0.545	5.28	113	9	56	24	62	54	1.82
KBO 리그 통산 (8시즌)		248	1414⅔	103	74	26	0.582	**2.46** (2위) [38]	1249	70	419	1019	478	387	1.15

출처 : 스탯티즈 (http://www.statiz.co.kr/main.php)

Data 구성

- 2015~2020 전체 투수 경기 스탯

- 1983~1988 최동원 선수 경기 스탯

- 2016~2021 연봉 총액 1억 이상 선수 명단

문제 1. 2015~2020 선수 경기 스탯 및 연봉 / 1983~1988 최동원 선수 스탯

왜 1983~1988 시즌 다른 투수들의 정보는 배제했나?

- 97년 이전의 연봉 데이터 부재

- 목표 : 2015~2020 활동 시 최동원 선수의 연봉을 예측하는 것이 목표

문제 1. 2015~2020 선수 경기 스탯 / 1983~1988 최동원 선수 스탯 수집

데이터 수집

- 활용 모듈
- 데이터 출처
- 수집 절차
- 데이터 맞춤

활용 모듈

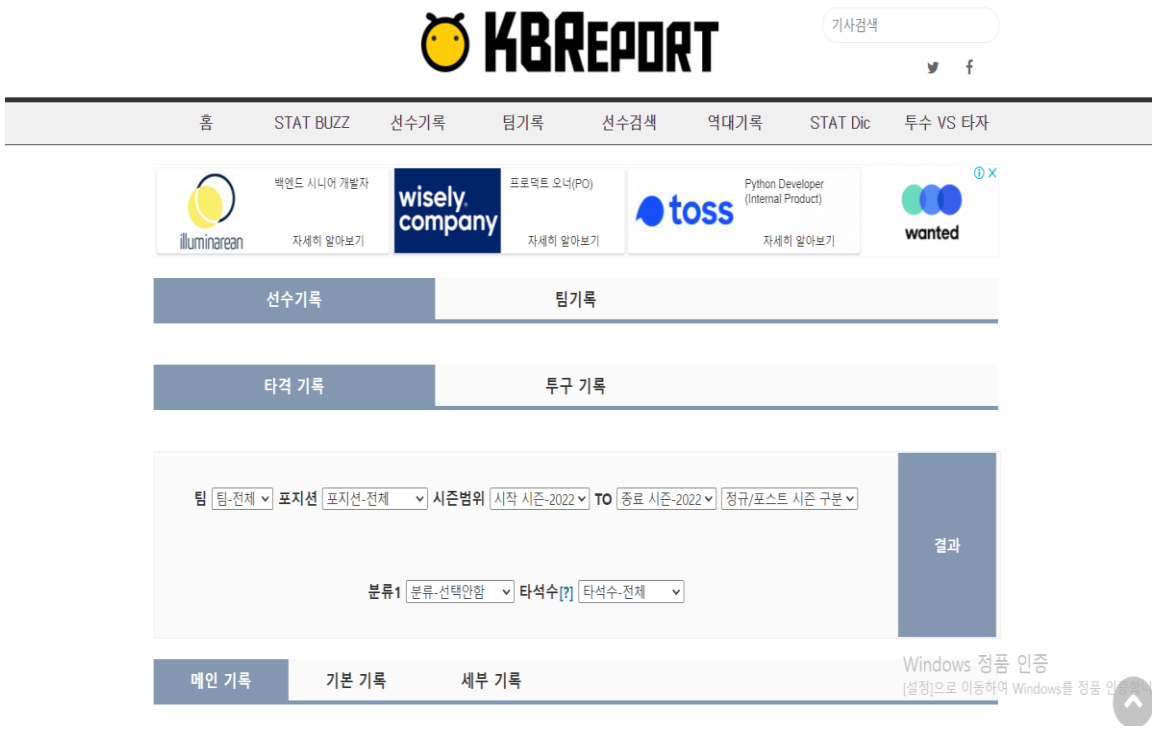
- BeautifulSoup
- Html_table_parser
- Pandas
- Selenium

활용 모듈

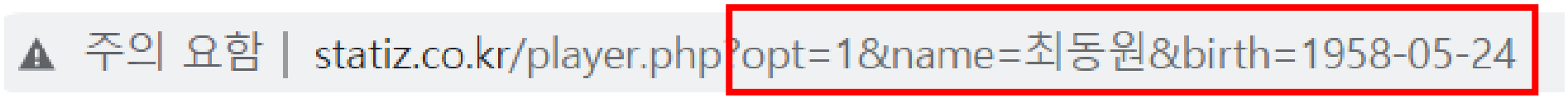
- BeautifulSoup
- -> 각 페이지에 기재된 정보 수집
- Html_table_parser
- -> 페이지에 기재된 테이블 데이터를 쉽게 수집하게 함
- Pandas
- -> 수집한 데이터 가공
- Selenium
- -> 페이지 이동

데이터 출처

- KBReport(http://www.kbreport.com/leader/pitcher/main)
- Statiz(http://www.statiz.co.kr/stat.php)



Url 활용 접속 -> 수집



[11]
```
choi.columns
```

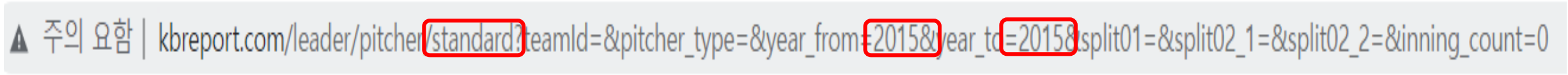
```
Index(['연도', '팀', '나이', '출장', '완투', '완봉', '선발', '승', '패', '세', '홀드', '이닝',
       '실점', '자책', '타자', '안타', '2타', '3타', '홈런', '볼넷', '고4', '사구', '삼진', '보크',
       '폭투', 'ERA', 'FIP', 'WHIP', 'ERA+', 'FIP+', 'WAR', 'WPA', 'K/9', 'BB/9',
       'K/BB', 'HR/9', 'K%', 'BB%', 'K-BB%', 'PFR', 'BABIP', 'LOB%', '타율',
       '출루율', '장타율', 'OPS', 'WHIP+', '투구', 'IP/G', 'P/G', 'P/IP', 'P/PA',
       'CYP'],
      dtype='object')
```

[13]
```
choi
```

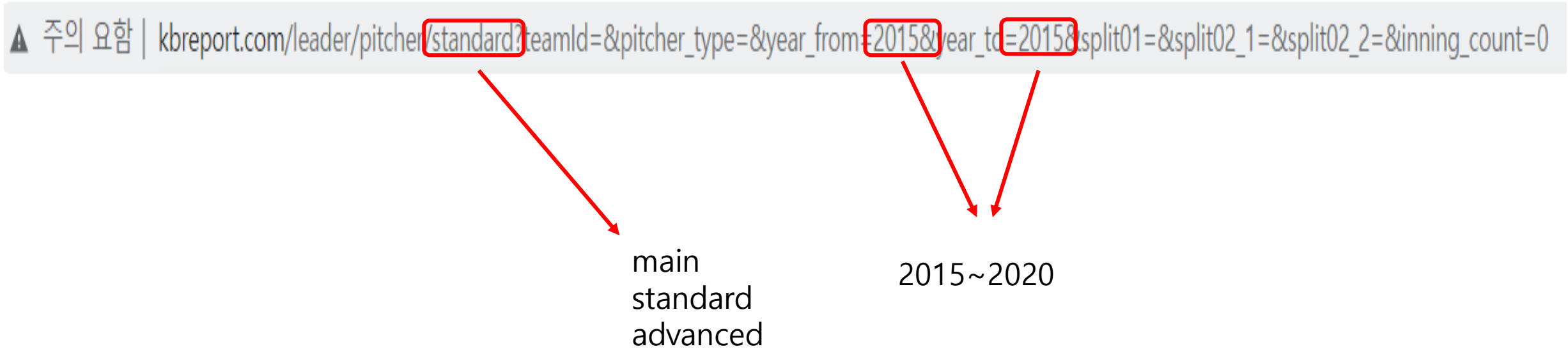
	연도	팀	나이	출장	완투	완봉	선발	승	패	세	...	출루율	장타율	OPS	WHIP+	투구	IP/G	P/G	P/IP	P/PA	CYP
0	1983	롯데	25	38	16	1	21	9	16	4	...	0.309			1.25		5.5				61.1
1	1984	롯데	26	51	14	1	20	27	13	6	...	0.280			1.09		5.6				115.6
2	1985	롯데	27	42	14	4	17	20	9	8	...	0.257			0.97		5.4				100.6
3	1986	롯데	28	39	17	4	21	19	14	2	...	0.262			0.99		6.9				127.3
4	1987	롯데	29	32	15	4	22	14	12	2	...	0.317			1.28		7.0				72.3
5	1988	롯데	30	16	3	1	4	7	3	3	...	0.298			1.22		5.2				38.0

6 rows × 53 columns

Url 활용 접속 -> 페이지 이동 -> 수집



Url 활용 접속 -> 페이지 이동 -> 수집



```
[15]
whole.columns
```
Python

```
Index(['#', '선수명', '팀명', '승', '패', '세', '홀드', '블론', '경기', '선발', '이닝', '삼진/9',
       '볼넷/9', '홈런/9', 'BABIP', 'LOB%', 'ERA', 'RA9-WAR', 'FIP', 'kFIP', 'WAR',
       '완투', '완봉', 'QS', '타자', '안타', '2루타', '3루타', '홈런', '실점', '자책', '삼진',
       '볼넷', '고4', 'HBP', '폭투', '보크', 'PK', '도루', '도실', '삼진%', '볼넷%', '삼/볼',
       '피안타율', '피출루율', '피장타율', '피OPS', 'WHIP', '연도'],
      dtype='object')
```

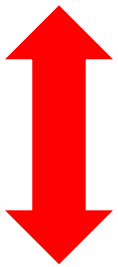
```
[16]
whole.head()
```
Python

	#	선수명	팀명	승	패	세	홀드	블론	경기	선발	...	도실	삼진%	볼넷%	삼/볼	피안타율	피출루율	피장타율	피OPS	WHIP	연도
0	1	소사	LG	10	12	0	1	0	32	30	...	11	21.9	4.4	4.92	0.266	0.302	0.402	0.704	1.21	2015
1	2	밴헤켄	Hero	15	8	0	0	0	32	32	...	5	23.4	8.1	2.88	0.257	0.318	0.371	0.689	1.31	2015
2	3	해커	NC	19	5	0	0	0	31	31	...	6	19.7	4.3	4.56	0.232	0.287	0.333	0.621	1.03	2015
3	4	윤성환	삼성	17	8	0	0	0	30	30	...	5	20.3	3.7	5.47	0.264	0.299	0.427	0.726	1.18	2015
4	5	린드블럼	롯데	13	11	0	0	0	32	32	...	11	20.9	6.0	3.46	0.250	0.306	0.406	0.711	1.18	2015

5 rows × 49 columns

데이터 맞춤

- 최동원 선수 스탯 데이터의 필드 수와 속성



- 2015~2020 전체 투수 스탯 데이터의 필드 수와 속성

수정(dtype) -> 삭제 -> 변경(필드명)

[40]
```
whole_copy.columns
```

```
Index(['선수명', '팀명', '경기', '완투', '완봉', '선발', '승', '패', '세', '홀드', '이닝', '실점',
       '자책', '타자', '안타', '홈런', '볼넷', '고4', 'HBP', '삼진', '보크', '폭투', 'ERA',
       'FIP', 'WHIP', 'WAR', '삼진/9', '볼넷/9', '삼/볼', '홈런/9', '삼진%', '볼넷%',
       'BABIP', 'LOB%', '피안타율', '피출루율', '연도'],
      dtype='object')
```

[41]
```
choi_copy.columns
```

```
Index(['팀명', '경기', '완투', '완봉', '선발', '승', '패', '세', '홀드', '이닝', '실점', '자책',
       '타자', '안타', '홈런', '볼넷', '고4', 'HBP', '삼진', '보크', '폭투', 'ERA', 'FIP',
       'WHIP', 'WAR', '삼진/9', '볼넷/9', '삼/볼', '홈런/9', '삼진%', '볼넷%', 'BABIP',
       'LOB%', '피안타율', '피출루율'],
      dtype='object')
```

2. 선발/마무리 투수의 구분은 그 해의 경기 수 대비 선발 출장 경기 수를 비교하여 결정한다.

- -> 규정이닝 선수 표본수의 문제

- -> ‘스윙맨 ' 보직의 존재

```
whole_rm_rest
```

Python

	경기	완투	완봉	선발	승	패	세	홀드	이닝	실점	...	볼넷%	BABIP	LOB%	피안타율	피출루율	연도	보직전	보직	선수명	팀명
0	32.0	2.0	1.0	30.0	10.0	12.0	0.0	1.0	194.1	102.0	...	4.4	0.327	63.4	0.266	0.302	2015.0	0.937500	1	소사	LG
1	32.0	0.0	0.0	32.0	15.0	8.0	0.0	0.0	196.2	92.0	...	8.1	0.328	69.6	0.257	0.318	2015.0	1.000000	1	밴헤켄	Hero
2	31.0	1.0	0.0	31.0	19.0	5.0	0.0	0.0	204.0	81.0	...	4.3	0.276	71.5	0.232	0.287	2015.0	1.000000	1	해커	NC
3	30.0	3.0	1.0	30.0	17.0	8.0	0.0	0.0	194.0	86.0	...	3.7	0.302	76.0	0.264	0.299	2015.0	1.000000	1	윤성환	삼성
4	32.0	2.0	1.0	32.0	13.0	11.0	0.0	0.0	210.0	86.0	...	6.0	0.290	78.8	0.250	0.306	2015.0	1.000000	1	린드블럼	롯데
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1568	20.0	0.0	0.0	10.0	3.0	3.0	0.0	0.0	44.1	30.0	...	13.0	0.271	71.9	0.269	0.379	2020.0	0.500000	1	조영건	Hero
1569	33.0	0.0	0.0	0.0	2.0	0.0	0.0	2.0	37.2	20.0	...	6.2	0.299	89.2	0.315	0.362	2020.0	0.000000	0	임규빈	Hero
1570	12.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	18.2	27.0	...	15.2	0.371	63.5	0.381	0.495	2020.0	0.000000	0	박진태	KIA
1571	14.0	0.0	0.0	13.0	2.0	4.0	0.0	0.0	60.2	45.0	...	12.6	0.271	69.1	0.283	0.379	2020.0	0.928571	1	최성영	NC
1572	27.0	0.0	0.0	0.0	1.0	2.0	1.0	1.0	25.2	22.0	...	13.1	0.294	63.4	0.291	0.397	2020.0	0.000000	0	이형범	두산

1573 rows × 39 columns

```
whole_rm_rest[whole_rm_rest['이닝']>=144]
```

[56] Python

	경기	완투	완봉	선발	승	패	세	홀드	이닝	실점	...	볼넷%	BABIP	LOB%	피안타율	피출루율	연도	보직전	보직	선수명	팀명
0	32.0	2.0	1.0	30.0	10.0	12.0	0.0	1.0	194.1	102.0	...	4.4	0.327	63.4	0.266	0.302	2015.0	0.937500	1	소사	LG
1	32.0	0.0	0.0	32.0	15.0	8.0	0.0	0.0	196.2	92.0	...	8.1	0.328	69.6	0.257	0.318	2015.0	1.000000	1	밴헤켄	Hero
2	31.0	1.0	0.0	31.0	19.0	5.0	0.0	0.0	204.0	81.0	...	4.3	0.276	71.5	0.232	0.287	2015.0	1.000000	1	해커	NC
3	30.0	3.0	1.0	30.0	17.0	8.0	0.0	0.0	194.0	86.0	...	3.7	0.302	76.0	0.264	0.299	2015.0	1.000000	1	윤성환	삼성
4	32.0	2.0	1.0	32.0	13.0	11.0	0.0	0.0	210.0	86.0	...	6.0	0.290	78.8	0.250	0.306	2015.0	1.000000	1	린드블럼	롯데
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1310	29.0	0.0	0.0	29.0	11.0	9.0	0.0	0.0	157.2	87.0	...	9.1	0.316	68.1	0.266	0.343	2020.0	1.000000	1	라이트	NC
1316	28.0	1.0	1.0	28.0	10.0	13.0	0.0	0.0	165.0	107.0	...	5.7	0.330	65.3	0.304	0.354	2020.0	1.000000	1	서폴드	한화
1317	29.0	0.0	0.0	28.0	13.0	11.0	0.0	0.0	157.1	90.0	...	11.3	0.299	68.9	0.249	0.357	2020.0	0.965517	1	박종훈	SK
1320	28.0	0.0	0.0	28.0	8.0	10.0	0.0	0.0	147.1	85.0	...	7.1	0.333	72.1	0.298	0.354	2020.0	1.000000	1	박세웅	롯데
1336	30.0	0.0	0.0	30.0	6.0	15.0	0.0	0.0	162.0	121.0	...	11.7	0.336	65.3	0.301	0.391	2020.0	1.000000	1	핀토	SK

128 rows × 39 columns

128명... 그마저도 대부분은 외국인 용병

2. 선발/마무리 투수의 구분은 그 해의 경기 수 대비 선발 출장 경기 수를 비교하여 결정한다.

- -> 규정이닝 선수 표본수의 문제

- -> '스윙맨 ' 보직의 존재

스윙맨

- 두 가지 이상의 포지션을 겸하는 선수

- 야구 종목에서는 선발과 중간계투 사이에서 전천후로 뛰며 윤활유 역할을 해주는 투수를 일컫는다.

사전 정의상 중간계투로 분류되므로,
스윙맨은 선발에서 제외

2. 2015~2020년 선수들과 연봉이 결정되는 해인 2016~2021년 연봉 데이터 수집 및 기존 데이터와 매핑

문제 2. 2015~2020년 선수들과 연봉이 결정되는 해인 2016~2021년 연봉 데이터 수집 및 기존 데이터와 매핑

데이터 매핑

- 배제 대상 선정
- 활용 데이터
- 매핑 절차

배제 대상 선정

- 고졸 및 대졸 신인 선수

- 외국인 용병

- 정규 시즌 데이터 기반 다음 해의 연봉 예측 모델

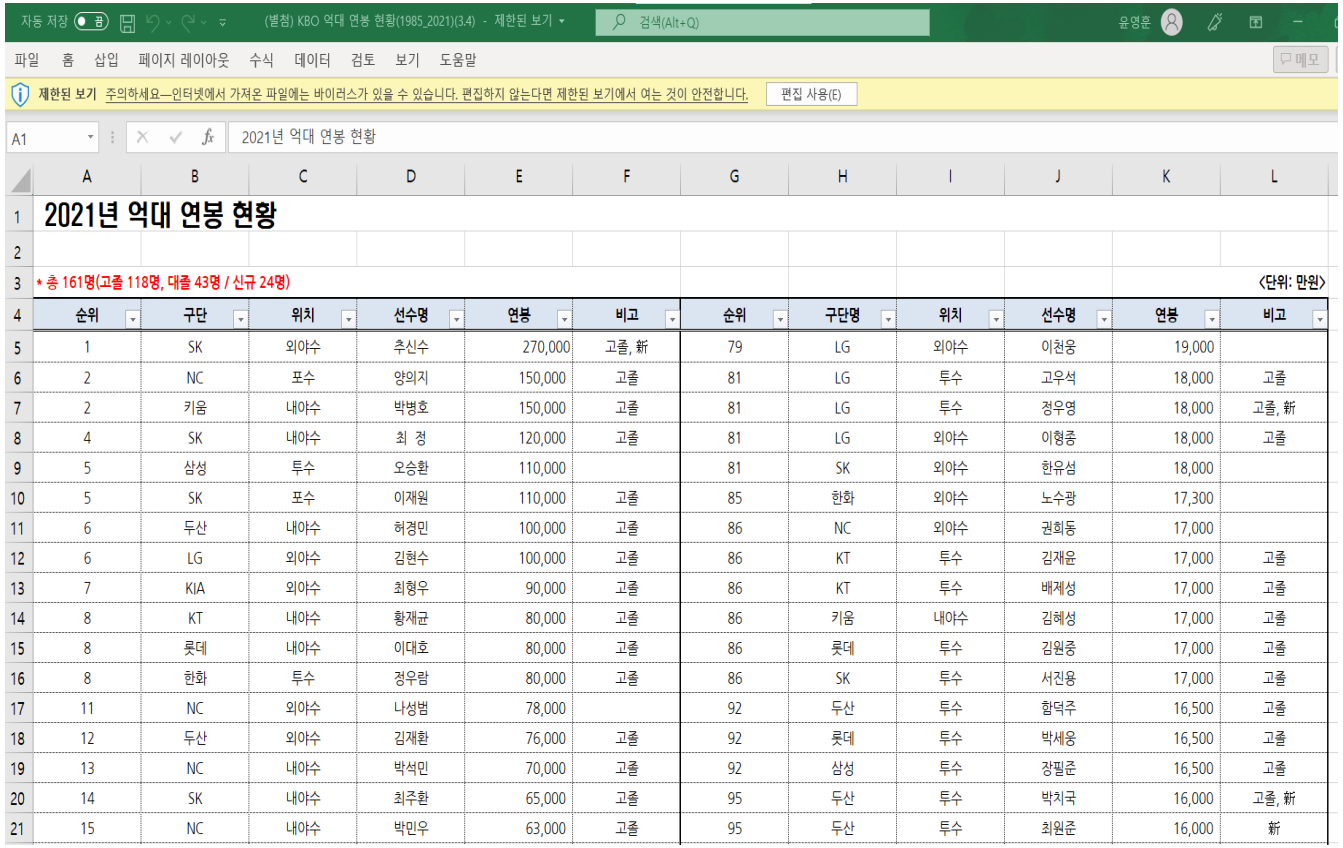
- 대상 선수들의 연봉 책정은 고등학교, 대학교 또는 타 리그에서의 성적을 기반으로 함

활용 데이터

(별첨) KBO 억대 연봉 현황 (1985_2021)(3.4).xlsx 파일

출처 : KBO 보도자료

(https://www.koreabaseball.com/News/Notice/View.aspx?bdSe=7980)



(별첨) KBO 억대 연봉 현황(1985_2021)(3.4) - 제한된 보기

제한된 보기 주의하세요—인터넷에서 가져온 파일에는 바이러스가 있을 수 있습니다. 편집하지 않는다면 제한된 보기에서 여는 것이 안전합니다. 편집 사용(E)

A1 2021년 억대 연봉 현황

2021년 억대 연봉 현황

* 총 161명(고졸 118명, 대졸 43명 / 신규 24명)

〈단위: 만원〉

순위	구단	위치	선수명	연봉	비고	순위	구단명	위치	선수명	연봉	비고
1	SK	외야수	추신수	270,000	고졸, 新	79	LG	외야수	이천웅	19,000	
2	NC	포수	양의지	150,000	고졸	81	LG	투수	고우석	18,000	고졸
2	키움	내야수	박병호	150,000	고졸	81	LG	투수	정우영	18,000	고졸, 新
4	SK	내야수	최 정	120,000	고졸	81	LG	외야수	이형종	18,000	고졸
5	삼성	투수	오승환	110,000		81	SK	외야수	한유섬	18,000	
5	SK	포수	이재원	110,000	고졸	85	한화	외야수	노수광	17,300	
6	두산	내야수	허경민	100,000	고졸	86	NC	외야수	권희동	17,000	
6	LG	외야수	김현수	100,000	고졸	86	KT	투수	김재윤	17,000	고졸
7	KIA	외야수	최형우	90,000	고졸	86	KT	투수	배제성	17,000	고졸
8	KT	내야수	황재균	80,000	고졸	86	키움	내야수	김혜성	17,000	고졸
8	롯데	내야수	이대호	80,000	고졸	86	롯데	투수	김원중	17,000	고졸
8	한화	투수	정우람	80,000	고졸	86	SK	투수	서진용	17,000	고졸
11	NC	외야수	나성범	78,000		92	두산	투수	함덕주	16,500	고졸
12	두산	외야수	김재환	76,000	고졸	92	롯데	투수	박세웅	16,500	고졸
13	NC	내야수	박석민	70,000	고졸	92	삼성	투수	장필준	16,500	고졸
14	SK	내야수	최주환	65,000	고졸	95	두산	투수	박치국	16,000	고졸, 新
15	NC	내야수	박민우	63,000	고졸	95	두산	투수	최원준	16,000	新

왜 억대 연봉자 데이터인가?

왜 억대 연봉자 데이터인가?

-> 연봉은 실력을 반영하는가?

- 최저 연봉자 표본 배제 : 실력 영향 증가

- KBO의 FA제도 : 실력에 비해 저평가

매핑 절차

- 연봉 데이터 수정

-> 타자 배제, 필드명 수정

- Merge 함수를 통한 연봉 데이터 매핑

- 특정 필드 배제

-> 연도 및 팀명

```
extra_analysis
```

[94]

	경기	완투	완봉	선발	승	패	세	홀드	이닝	실점	...	삼/볼	홈런/9	삼진%	볼넷%	BABIP	LOB%	피안타율	피출루율	보직	연봉
0	30.0	3.0	1.0	30.0	17.0	8.0	0.0	0.0	194.0	86.0	...	5.47	1.25	20.3	3.7	0.302	76.0	0.264	0.299	1	80000.0
1	25.0	1.0	0.0	25.0	11.0	9.0	0.0	0.0	152.2	64.0	...	7.00	0.77	19.2	2.7	0.333	73.6	0.282	0.314	1	40000.0
2	30.0	1.0	1.0	29.0	14.0	6.0	0.0	1.0	176.2	86.0	...	2.42	0.97	21.2	8.8	0.311	72.4	0.257	0.326	1	85000.0
3	32.0	1.0	1.0	31.0	15.0	6.0	0.0	1.0	184.1	52.0	...	2.01	0.88	20.8	10.3	0.277	87.2	0.232	0.319	1	75000.0
4	30.0	1.0	1.0	30.0	18.0	5.0	0.0	0.0	189.2	84.0	...	2.86	1.09	16.0	5.6	0.296	75.2	0.269	0.312	1	40000.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
395	28.0	0.0	0.0	9.0	3.0	7.0	0.0	1.0	77.2	49.0	...	1.42	1.04	10.6	7.4	0.317	68.2	0.304	0.364	0	15500.0
396	33.0	0.0	0.0	8.0	1.0	6.0	0.0	4.0	62.0	52.0	...	1.10	1.16	15.3	13.9	0.283	56.0	0.260	0.370	0	16000.0
397	23.0	0.0	0.0	0.0	2.0	2.0	0.0	3.0	20.1	20.0	...	1.17	1.33	21.4	18.4	0.288	53.3	0.240	0.396	0	28000.0
398	63.0	0.0	0.0	0.0	2.0	7.0	8.0	12.0	61.0	33.0	...	1.65	1.62	21.2	12.9	0.259	75.4	0.234	0.335	0	17000.0
399	37.0	0.0	0.0	3.0	4.0	4.0	0.0	6.0	41.0	52.0	...	1.18	1.76	12.6	10.6	0.384	47.6	0.370	0.434	0	10500.0

400 rows × 36 columns

문제 3. 최동원 선수 연봉 예측 (각 시즌)

연봉 예측

- 데이터 탐색 및 분석
- 결정 변수 설정 및 변수 설명
- 모델 설정
- 검증
- 예측
- 결론

데이터 탐색 및 분석

- 탐색 및 분석 과정은 거쳐간 실습 내용 중 ‘보스턴 집값 예측’ 참고

데이터 : 수치 확인

```
...              경기        완투        완봉        선발         승          패  \
count  400.000000  400.000000  400.000000  400.000000  400.000000  400.000000
mean    42.520000    0.115000    0.047500    7.597500    4.847500    4.472500
std     18.318393    0.414977    0.235334   10.833925    3.838108    3.135894
min      1.000000    0.000000    0.000000    0.000000    0.000000    0.000000
25%     28.000000    0.000000    0.000000    0.000000    2.000000    2.000000
50%     43.000000    0.000000    0.000000    0.000000    4.000000    4.000000
75%     59.000000    0.000000    0.000000   17.000000    7.000000    6.000000
max     78.000000    3.000000    2.000000   31.000000   20.000000   14.000000

              세         홀드         이닝         실점
count  400.000000  400.000000  400.000000  400.00000
mean     3.847500    5.455000   76.688250   41.96000
std      7.839188    6.805053   44.802491   25.73381
min      0.000000    0.000000    0.100000    2.00000
25%      0.000000    0.000000   47.200000   23.75000
50%      0.000000    3.000000   62.000000   34.00000
75%      3.000000    9.000000  107.250000   62.00000
max     37.000000   40.000000  200.100000  119.00000
```

데이터 : 수치 확인

	자책	타자	안타	홈런	볼넷	고4 \
count	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000
mean	38.447500	335.910000	81.490000	8.515000	27.465000	1.222500
std	23.802703	193.507091	49.108288	6.215009	17.152823	1.497866
min	2.000000	4.000000	3.000000	0.000000	0.000000	0.000000
25%	21.750000	206.500000	49.000000	4.000000	15.000000	0.000000
50%	32.000000	270.500000	64.000000	7.000000	23.000000	1.000000
75%	56.000000	479.750000	123.000000	12.000000	37.000000	2.000000
max	115.000000	850.000000	228.000000	29.000000	91.000000	7.000000

	HBP	삼진	보크	폭투
count	400.000000	400.000000	400.000000	400.000000
mean	4.540000	62.245000	0.167500	3.492500
std	4.305199	35.646904	0.430007	3.116766
min	0.000000	1.000000	0.000000	0.000000
25%	2.000000	37.750000	0.000000	1.000000
50%	3.500000	55.000000	0.000000	3.000000
75%	6.000000	85.000000	0.000000	5.000000
max	25.000000	194.000000	3.000000	17.000000

데이터 : 수치 확인

…	ERA	FIP	WHIP	WAR	삼진/9	볼넷/9	\
count	400.000000	400.000000	400.000000	400.000000	400.000000	400.000000	
mean	4.811700	4.690575	1.473975	1.069100	7.481100	3.388275	
std	3.069098	1.054579	0.492286	1.153844	2.054254	1.457845	
min	1.040000	-2.600000	0.670000	-0.860000	2.080000	0.000000	
25%	3.610000	4.045000	1.290000	0.230000	6.190000	2.522500	
50%	4.530000	4.650000	1.430000	0.775000	7.360000	3.210000	
75%	5.370000	5.340000	1.570000	1.495000	8.585000	4.090000	
max	54.000000	8.400000	9.000000	7.430000	27.000000	13.500000	

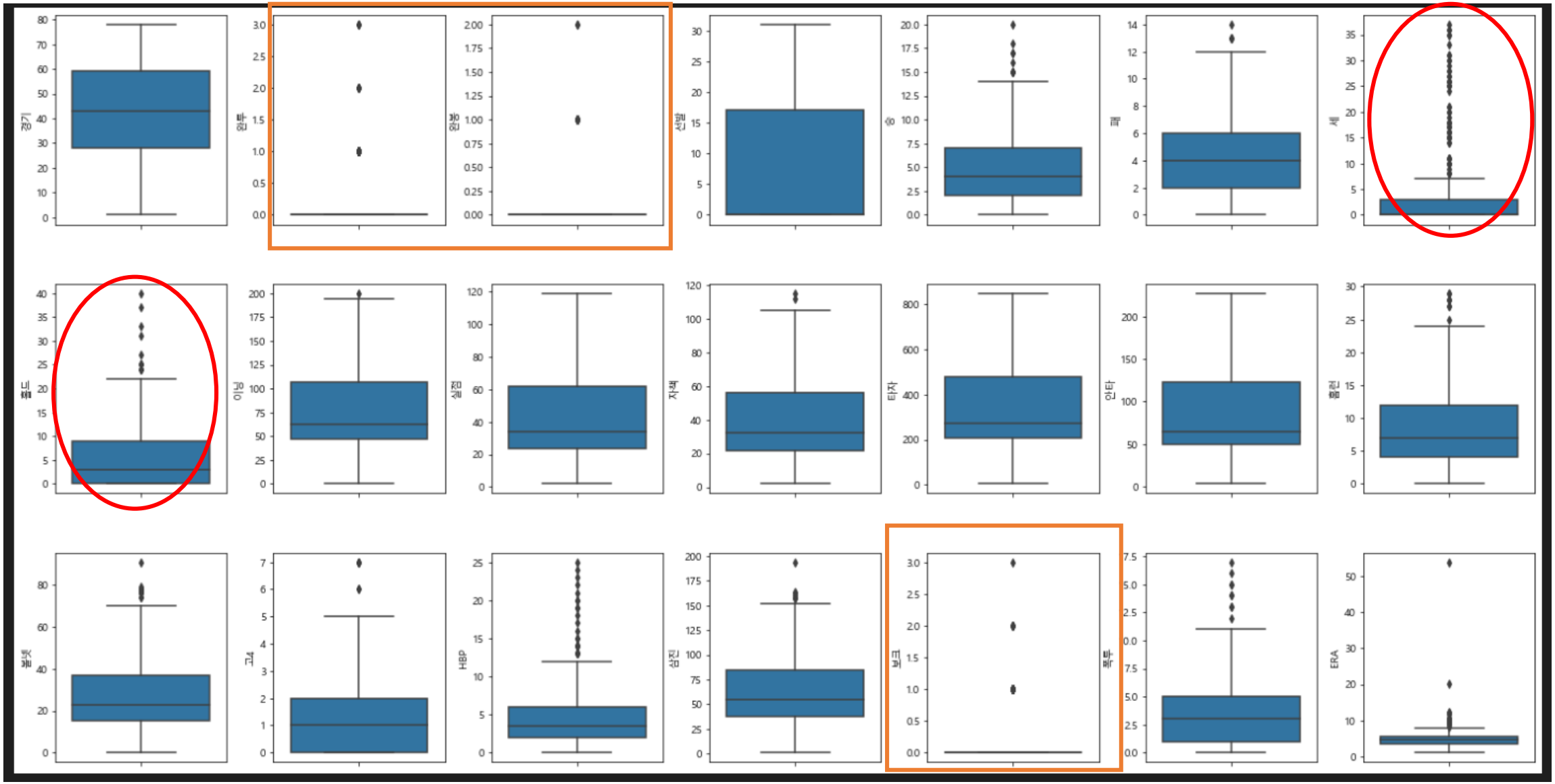
	삼/볼	홈런/9	삼진%	볼넷%
count	400.000000	400.000000	400.00000	400.00000
mean	2.488875	1.004700	18.85675	8.42825
std	1.170134	0.506178	4.84437	3.06454
min	0.000000	0.000000	4.50000	0.00000
25%	1.687500	0.640000	15.67500	6.40000
50%	2.255000	0.980000	18.70000	8.10000
75%	2.952500	1.290000	21.80000	10.30000
max	7.810000	2.950000	32.50000	21.40000

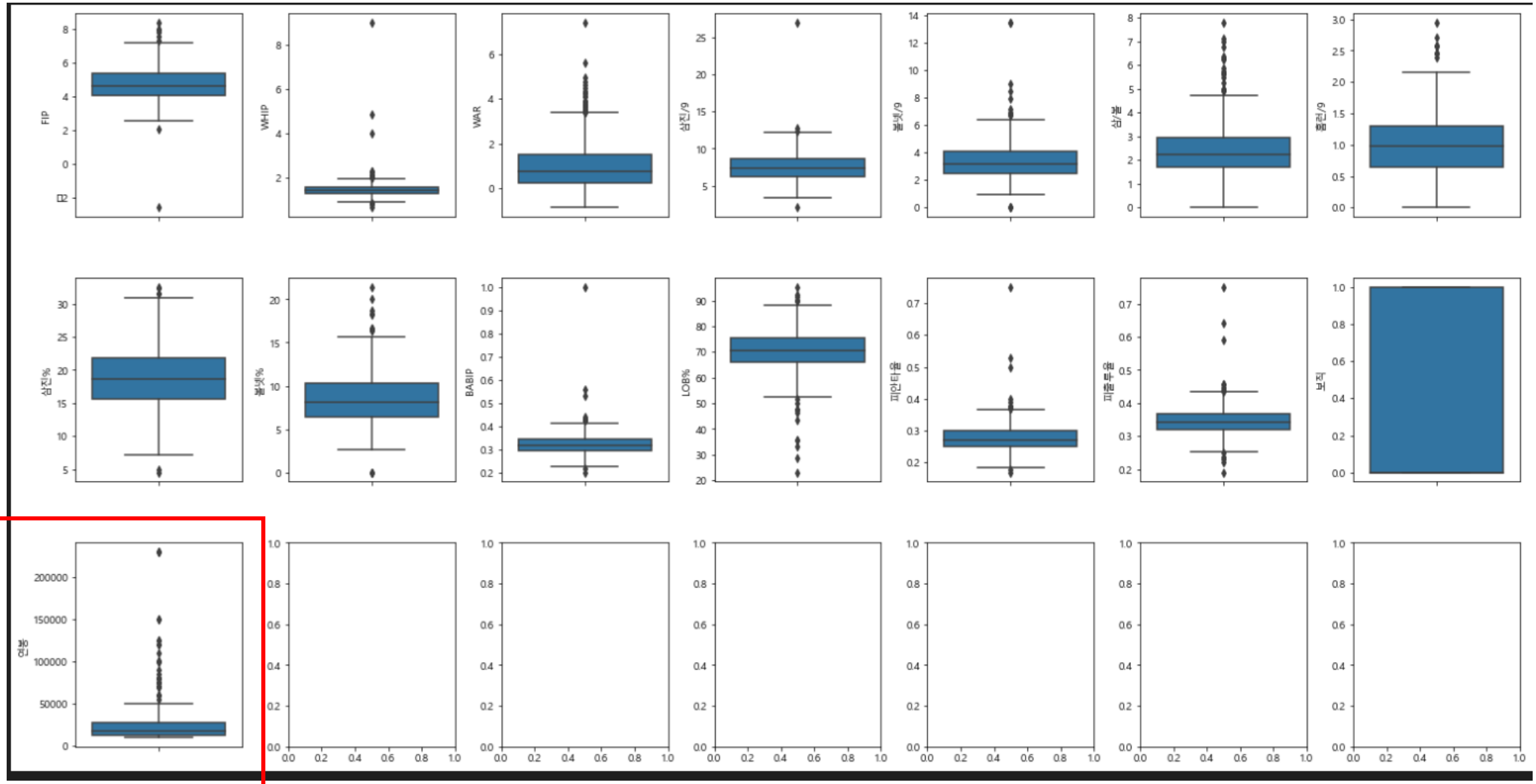
데이터 : 수치 확인

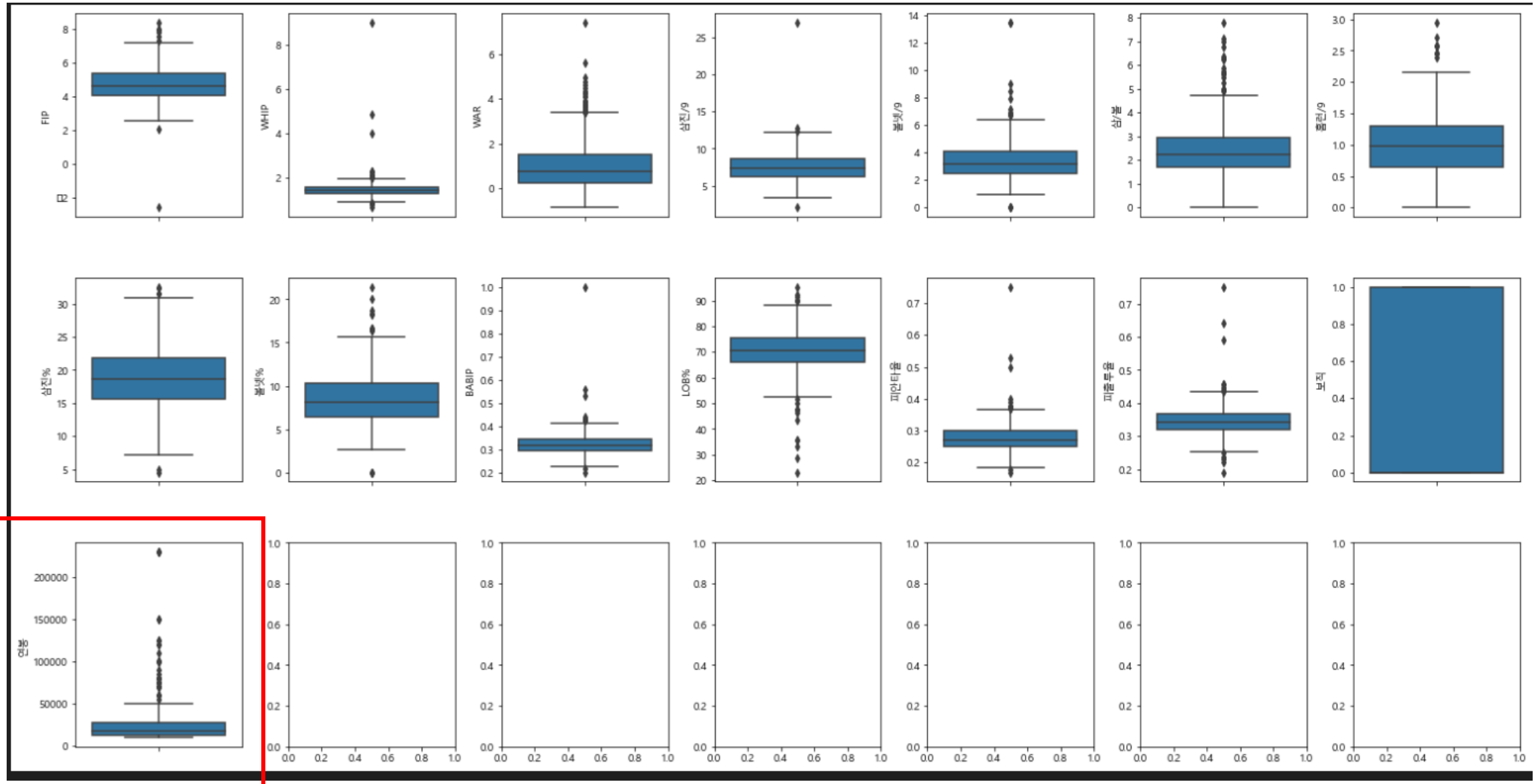
...	BABIP	LOB%	피안타율	피출루율	보직 \
count	400.000000	400.000000	400.000000	400.000000	400.000000
mean	0.322500	70.126000	0.276338	0.345242	0.277500
std	0.053876	8.798768	0.047462	0.046655	0.448326
min	0.200000	23.100000	0.169000	0.190000	0.000000
25%	0.295750	66.000000	0.252000	0.320000	0.000000
50%	0.318000	70.750000	0.271500	0.343000	0.000000
75%	0.344000	75.525000	0.299000	0.366000	1.000000
max	1.000000	95.200000	0.750000	0.750000	1.000000

	연봉
count	400.000000
mean	27073.000000
std	28898.663514
min	10000.000000
25%	12500.000000
50%	17000.000000
75%	28000.000000
max	230000.000000

데이터 : 분포 확인


경기
완투
완봉
선발
승
패
세
홀드
이닝
실점
자책
타자
안타
홈런
볼넷
고4
HBP
삼진
보크
폭투
ERA


FIP
WHIP
WAR
삼진/9
볼넷/9
삼/볼
홈런/9
삼진%
볼넷%
BABIP
LOB%
피안타율
피홈루율
보직
연봉



연봉 수치에서 아웃라이어가 다수 확인됨

데이터 : 아웃라이어 확인

```
Column 피안타율 outliers = 3.50%
Column 피출루율 outliers = 3.25%
Column 보직 outliers = 0.00%
Column 연봉 outliers = 9.50%
```

데이터 : 아웃라이어 배제

```
[130]
```

```
extra_analysis_pro
```

	경기	완투	완봉	선발	승	패	세	홀드	이닝	실점	...	삼/볼	홈런/9	삼진%	볼넷%	BABIP	LOB%	피안타율	피출루율	보직	연봉
1	25.0	1.0	0.0	25.0	11.0	9.0	0.0	0.0	152.2	64.0	...	7.00	0.77	19.2	2.7	0.333	73.6	0.282	0.314	1	40000.0
4	30.0	1.0	1.0	30.0	18.0	5.0	0.0	0.0	189.2	84.0	...	2.86	1.09	16.0	5.6	0.296	75.2	0.269	0.312	1	40000.0
5	31.0	0.0	0.0	29.0	13.0	7.0	0.0	1.0	173.0	98.0	...	2.62	1.46	26.1	10.0	0.304	70.6	0.245	0.325	1	40000.0
8	45.0	0.0	0.0	17.0	11.0	4.0	0.0	10.0	123.1	70.0	...	2.98	1.02	22.5	7.6	0.327	68.6	0.265	0.338	0	30000.0
9	25.0	0.0	0.0	23.0	8.0	7.0	0.0	0.0	125.0	71.0	...	2.43	1.22	20.3	8.3	0.312	70.0	0.270	0.333	1	40000.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
395	28.0	0.0	0.0	9.0	3.0	7.0	0.0	1.0	77.2	49.0	...	1.42	1.04	10.6	7.4	0.317	68.2	0.304	0.364	0	15500.0
396	33.0	0.0	0.0	8.0	1.0	6.0	0.0	4.0	62.0	52.0	...	1.10	1.16	15.3	13.9	0.283	56.0	0.260	0.370	0	16000.0
397	23.0	0.0	0.0	0.0	2.0	2.0	0.0	3.0	20.1	20.0	...	1.17	1.33	21.4	18.4	0.288	53.3	0.240	0.396	0	28000.0
398	63.0	0.0	0.0	0.0	2.0	7.0	8.0	12.0	61.0	33.0	...	1.65	1.62	21.2	12.9	0.259	75.4	0.234	0.335	0	17000.0
399	37.0	0.0	0.0	3.0	4.0	4.0	0.0	6.0	41.0	52.0	...	1.18	1.76	12.6	10.6	0.384	47.6	0.370	0.434	0	10500.0

362 rows × 36 columns

남은 데이터는 362개

결정 변수 설정 및 변수 설명

결정 변수 파악 : 상관관계 분석

연봉	1.000000
안타	0.250882
이닝	0.242746
타자	0.241257
선발	0.231822
WAR	0.225506
승	0.223845
보직	0.219492
실점	0.206685
자책	0.206315
완투	0.203232
삼진	0.180163
패	0.175070
홈런	0.165670
볼넷	0.141539
HBP	0.096381
세	0.086960
완봉	0.079596
폭투	0.072914
보크	0.068994

결정 변수 파악 : 상관관계 분석

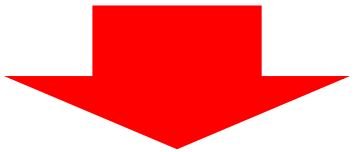
상관관계가
0.1이 넘는 변수들을
결정 변수로 설정

연봉	1.000000
안타	0.250882
이닝	0.242746
타자	0.241257
선발	0.231822
WAR	0.225506
승	0.223845
보직	0.219492
실점	0.206685
자책	0.206315
완투	0.203232
삼진	0.180163
패	0.175070
홈런	0.165670
볼넷	0.141539
HBP	0.096381
세	0.086960
완봉	0.079596
폭투	0.072914
보크	0.068994

결정 변수 파악 : 변수 설명

- 안타 : 한 시즌 동안 투수의 피안타의 개수
- 패 : 한 시즌 동안 투수의 총 패배
- 홈런 : 한 시즌 동안 투수의 총 피홈런의 수
- 실점 : 한 시즌 동안 투수의 총 실점
- 자책 : 한 시즌 동안 투수의 총 자책점(투수가 책임을 져야 할 실점)

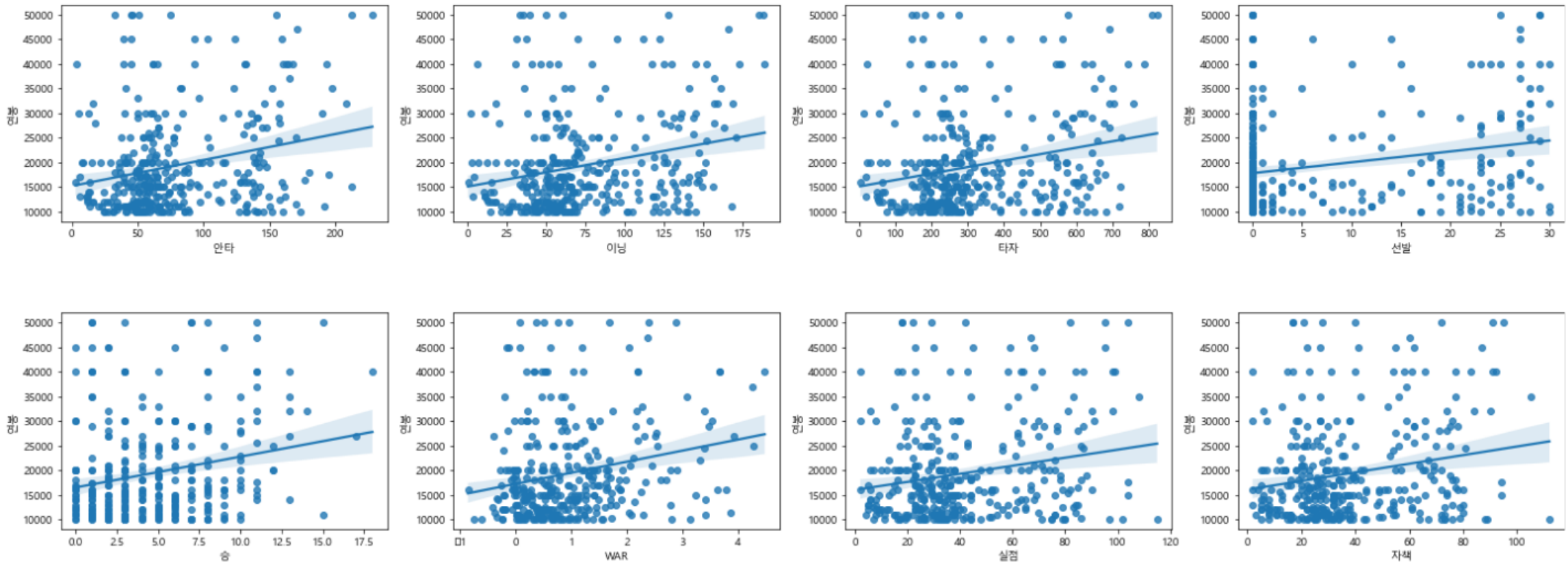
- 안타 : 한 시즌 동안 투수의 피안타의 개수
- 패 : 한 시즌 동안 투수의 총 패배
- 홈런 : 한 시즌 동안 투수의 총 피홈런의 수
- 실점 : 한 시즌 동안 투수의 총 실점
- 자책 : 한 시즌 동안 투수의 총 자책점(투수가 책임을 져야 할 실점)

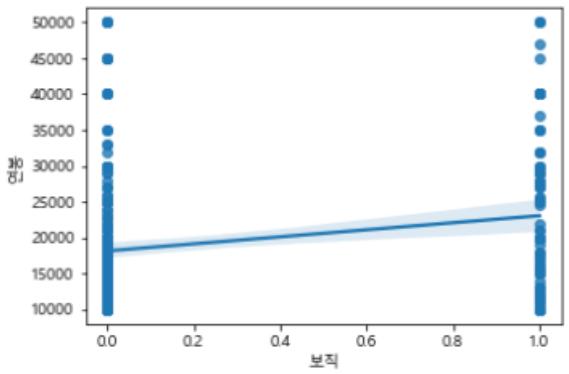


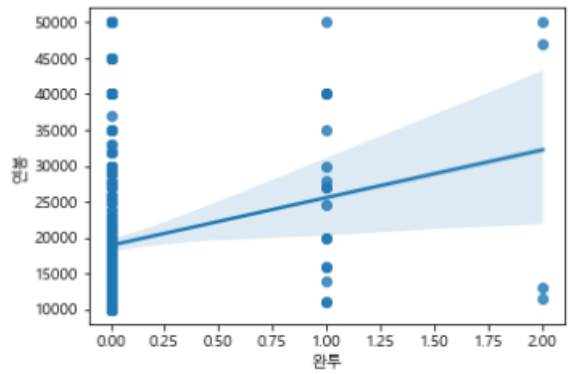
공통점 : 평균적으로 투수가 소화하는 이닝이 늘어날수록 같이 늘어나는 수치들

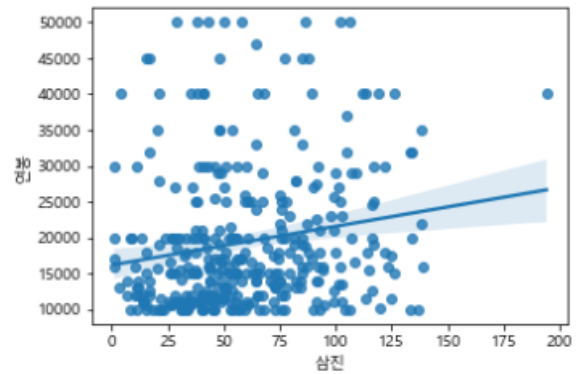
- 이닝 : 한 시즌 동안 투수가 던진 총 이닝 수
- 타자 : 한 시즌 동안 투수가 상대한 타자의 수
- 선발 : 한 시즌 동안 투수가 선발로 나온 경기의 수
- 승 : 한 시즌 동안 투수가 승리한 수
- 완투 : 한 시즌 동안 투수가 한 경기에서 시작부터 마지막까지 마운드를 지킨 횟수
- 삼진 : 한 시즌 동안 투수의 총 삼진 개수
- WAR : 대체선수 대비 승리기여도. 선수가 팀 승리에 얼마나 기여했는가를 표현하는 종합적인 성격의 스탯이다.
- 보직 : 한 시즌 동안 선발 수/경기 수
- 볼넷 : 한 시즌 동안 투수의 총 볼넷 개수

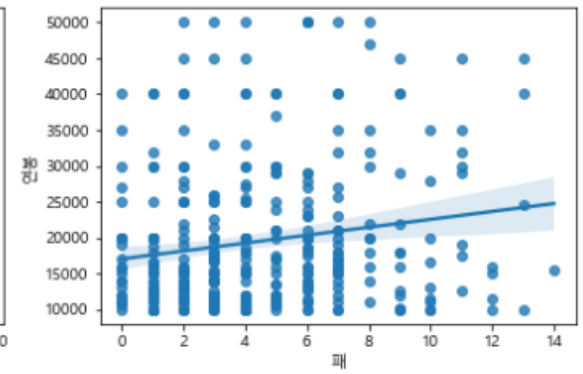
결정 변수 파악 : 변수별 분석

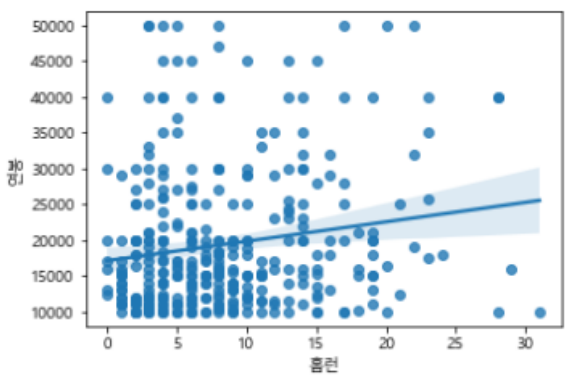

연봉
안타
이닝
타자
선발
승
WAR
실점
자책

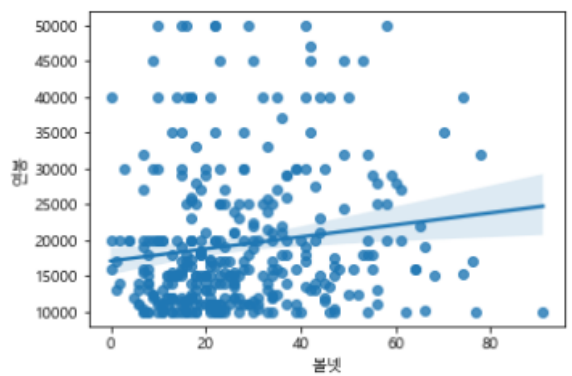

연봉
보직

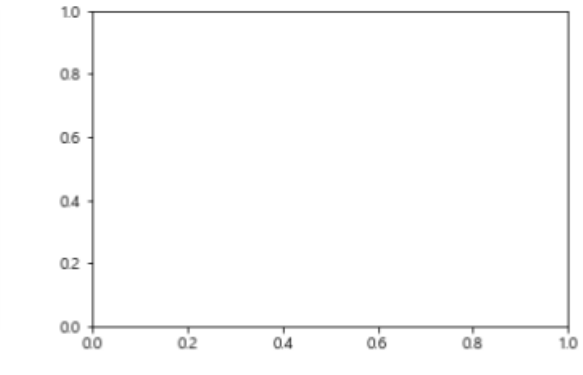

연봉
완투

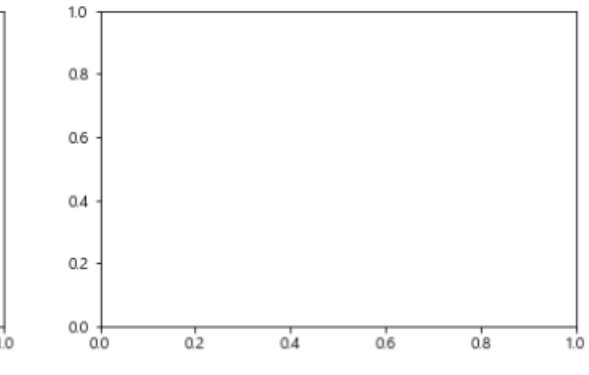

연봉
삼진


연봉
패


연봉
홈런


연봉
볼넷





OLS 분석 적용 결과

OLS Regression Results

Dep. Variable:	연봉	R-squared (uncentered):	0.762
Model:	OLS	Adj. R-squared (uncentered):	0.753
Method:	Least Squares	F-statistic:	77.48
Date:	Mon, 02 May 2022	Prob (F-statistic):	2.33e-96
Time:	14:45:47	Log-Likelihood:	-3761.2
No. Observations:	352	AIC:	7550.
Df Residuals:	338	BIC:	7605.
Df Model:	14		
Covariance Type:	nonrobust		

모델 설정

- Train_test_split 활용 : 위에서 분석한 개별 변수들을 X에, 연봉 칼럼을 y로 두고 진행

- Linear regression 모델 활용

-> 주택 가격 예측 모형과 같은 모델 활용

검증 : 평균 제곱근 편차(RMSE)

```
import numpy as np
from sklearn.metrics import mean_squared_error

pred_tr = reg.predict(X_train)
pred_test = reg.predict(X_test)
rmse_tr = (np.sqrt(mean_squared_error(y_train, pred_tr)))
rmse_test = (np.sqrt(mean_squared_error(y_test, pred_test)))

print(rmse_tr)
print(rmse_test)
```

✓ 0.7s

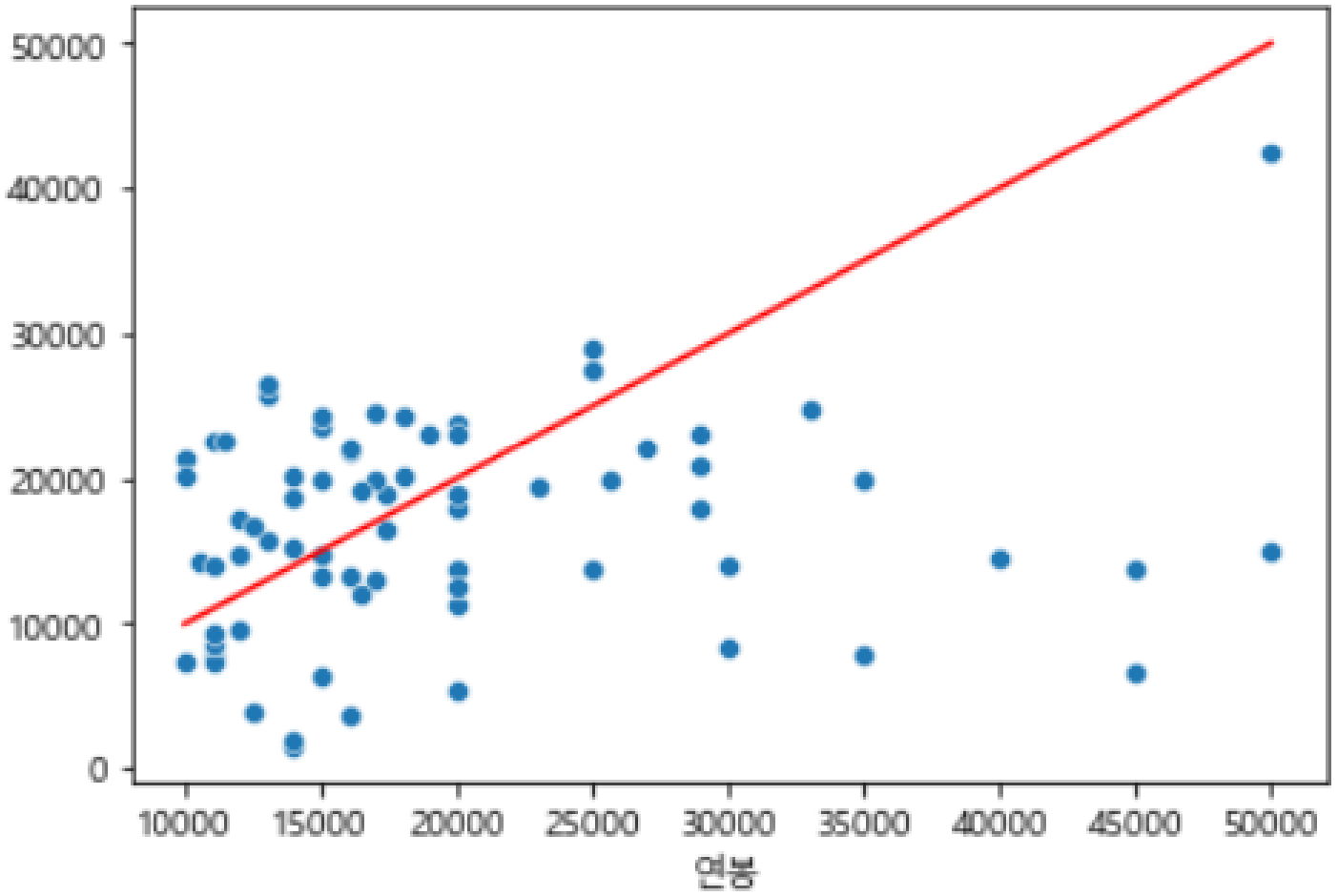
```
9217.126330039519
9679.38932751995
```

검증 : 평균 제곱근 편차(RMSE)

- MSE에 root를 씌움으로써, 큰 차이의 오차에 덜 민감

- 따라서 동일한 계산 단위를 적용하는 MAE나 에러값이 큰 MSE 대신 RMSE를 사용함

검증 :
예측값과
실제값
비교



예측 : 미리 예상해보기

최동원

안타	183.333333
이닝	215.250000
타자	861.333333
선발	17.500000
승	16.000000
WAR	7.891667
실점	67.333333
자책	54.333333
보직	0.500000
완투	13.166667
삼진	164.333333
패	11.166667
홈런	9.833333
볼넷	50.000000

dtype: float64

1억 이상 연봉자 평균

안타	82.708440
이닝	77.858568
타자	340.966752
선발	7.828645
승	4.943734
WAR	1.087980
실점	42.618926
자책	39.074169
보직	0.286445
완투	0.117647
삼진	63.122762
패	4.524297
홈런	8.670077
볼넷	27.818414

dtype: float64

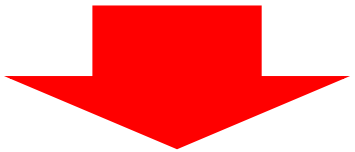
예측 : 분석

```
1983시즌 기반 다음해 연봉(예상) :  62806.14446121041
1984시즌 기반 다음해 연봉(예상) :  71047.67274198437
1985시즌 기반 다음해 연봉(예상) :  58446.84486036451
1986시즌 기반 다음해 연봉(예상) :  81156.11779385165
1987시즌 기반 다음해 연봉(예상) :  65793.2374087714
1988시즌 기반 다음해 연봉(예상) :  26807.44521116387
```

예측 : 분석

	안타	이닝	타자	선발	승	WAR	실점	자책	보직	완투	삼진	패	홈런	볼넷
0	202.0	208.2	863.0	21.0	9.0	5.00	89.0	67.0	1	16.0	148.0	16.0	17.0	51.0
84시즌 1	229.0	284.2	1132.0	20.0	27.0	9.72	91.0	76.0	0	14.0	223.0	13.0	18.0	68.0
2	170.0	225.0	865.0	17.0	20.0	9.88	60.0	48.0	0	14.0	161.0	9.0	7.0	41.0
86시즌 3	204.0	267.0	1039.0	21.0	19.0	11.74	60.0	46.0	1	17.0	208.0	14.0	7.0	55.0
4	218.0	224.0	920.0	22.0	14.0	7.15	80.0	70.0	1	15.0	163.0	12.0	6.0	61.0
5	77.0	83.1	349.0	4.0	7.0	3.86	24.0	19.0	0	3.0	83.0	3.0	4.0	24.0

- 안타 : 한 시즌 동안 투수의 피안타의 개수
- 패 : 한 시즌 동안 투수의 총 패배
- 홈런 : 한 시즌 동안 투수의 총 피홈런의 수
- 실점 : 한 시즌 동안 투수의 총 실점
- 자책 : 한 시즌 동안 투수의 총 자책점(투수가 책임을 져야 할 실점)



공통점 : 평균적으로 투수가 소화하는 이닝이 늘어날수록 같이 늘어나는 수치들

결론 : 최동원은 지금 기준으로도 어마어마한 선수

- 아웃라이어인 5억 5천의 연봉 수치를 가볍게 넘는 시즌이 6시즌 중 5시즌
- 보직이 선발인 경우, 평균적으로 더 많은 연봉을 받는다. (더 많은 이닝)
- WAR 수치로 인해서 부정적 지표가 어느 정도 해소됨

아쉬움

WAR 산출에서의 한계점

- 1. 조정 실점의 산출
- 2. 1승에 해당하는 점수(R/W)의 산출
- 3. 기대승률의 산출
- 4. 대체선수수준
- 5. 선발투수의 WAR

WAR 산출에서의 한계점

- 1. 조정 실점의 산출 -> **리그평균자책점, 리그평균실점**
- 2. 1승에 해당하는 점수(R/W)의 산출 -> **1승에 해당하는 점수 (어느 팀에 있느냐에 따라 다름)**
- 3. 기대승률의 산출 -> **구장 및 리그 평균의 타선, 불펜, 수비진 고려**
- 4. 대체선수수준 -> **해당 리그의 실정에 맞는 대체선수수준**
- 5. 선발투수의 WAR

- 혹시 코드에 대한 설명이 발표 과정에서 필요한지 궁금합니다.

- 부적절한 용어 사용 및 모델 결정과 스케일링 과정에서의 문제점 등 피드백 주시면 매우 감사하겠습니다. 바로 고치도록 노력하겠습니다!