KBO 데이터셋을 분석해보자

팀명 : 머니볼 (강민호, 박흥선, 함태식)

목차

1.
개요

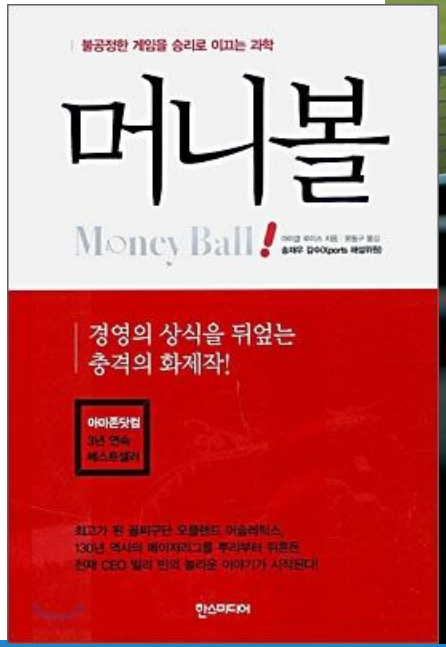
팀 명의 유래와 KBO 분석 프로젝트

팀 명 유래 (머니볼)

머니볼 : 현대 야구학의 경영 방법론

-> 약소 구단을 운영하기 위해선 데이터 분석을 통해 기량이 높은 선수를 싸게 영입 해야 한다.

- 영화화 되었음.



KBO 데이터셋 분석

- 1998 ~ 2018년도의 KBO 타자 데이터셋
- 1. 야구선수의 기량 성장 추이 분석
- 2. 야구선수의 기량에 따른 연봉 적합도 분석

위 두 가지 분석을 진행하는 것이 목표이다.

2.

KBO 데이터 소개

KBO 데이터 셋의 37개 컬럼 설명

KBO 데이터셋 소개

index	Column Name	설명
1	batter_name	선수 이름
2	age	나이
3	G	경기 수
4	PA	타수
5	AB	타석 수
6	R	득점
7	H	안타
8	2B	2루타
9	3B	3루타
10	HR	홈런

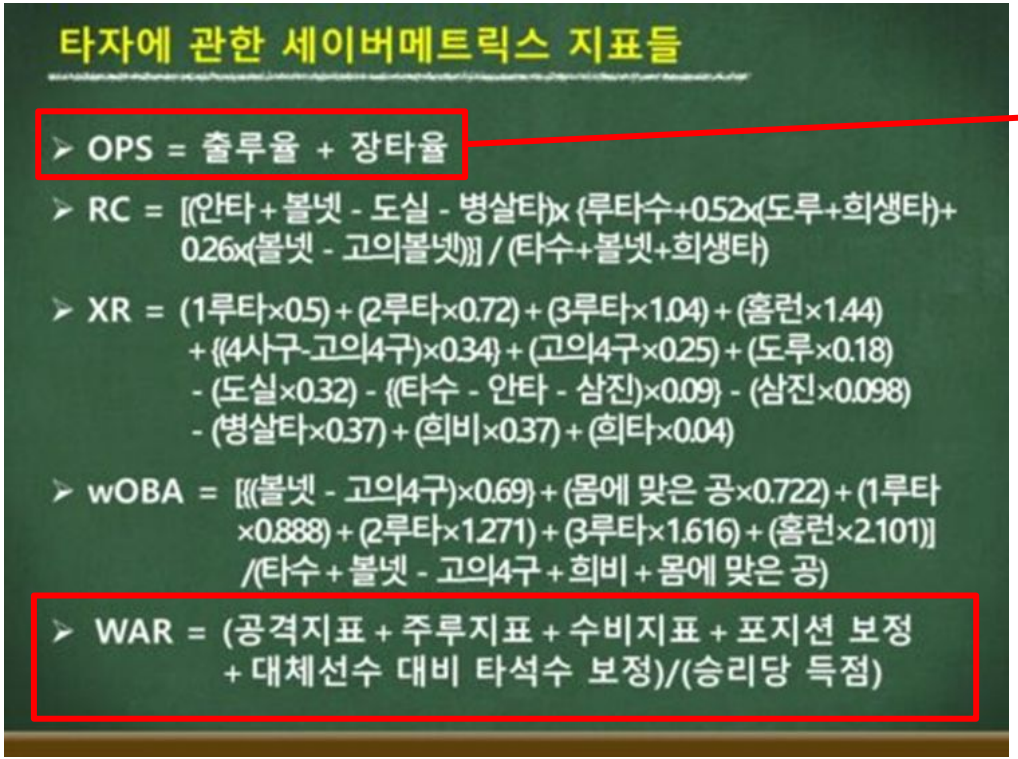
index	Column Name	설명
11	TB	총 루타 수
12	RBI	타점
13	SB	도루 성공
14	CS	도루 실패
15	BB	볼넷 수
16	HBP	몸에 맞은 공
17	GB	고의4구
18	SO	삼진
19	GDP	병살
20	BU	희생타

index	Column Name	설명
21	fly	희생 플라이
22	year	해당 시즌
23	salary	해당 시즌의 연봉
24	war	대체 선수 대비 승리 기여도
25	year_born	선수 태어난 연도
26	hand2	타석 위치
27	cp	최근 포지션
28	tp	통합 포지션
29	1B	1루타
30	FBP	BB + HBP

index	Column Name	설명
31	avg	타율
32	OBP	출루율
33	SLG	장타율
34	OPS	OBP + SLG
35	p_year	다음 시즌
36	YAB	다음 시즌 타석 수
37	YOPS	다음 시즌 OPS

***병살** : 미스플레이 없이 연속으로 아웃 카운트가 두 개 생기는 경우를 뜻하며, 그러한 수비에 참여한 수비수들에게 주어지는 기록 *희생타 : 번트 **희생 플라이 : 타자가 외야 쪽으로 플라이를 때렸을 때 수비가 그 공을 잡아서 아웃시키고 난 후, 3루에 있던 주자가 홈으로 출발하여, 수비측의 주자 태그보다 먼저 홈에 들어와서 득점한 것

추가 컬럼



- GPA(총생산평균, Gross Production Average) :
  = 1.8 * 출루율 + 장타율
- 출루율의 최댓값은 1, 장타율의 최댓값은 4이기 때문에 출루율을 보정한 상위 지표

- WAR(대체선수 대비 승리기여도, Wins Above Replacement) : 리그에 평균적인 대체선수에 비해 팀 승리에 얼마나 기여했는지 보여주는 지표 (모든 지표를 고려함)

3.
데이터 전처리

데이터 로드


```
In [2]: data_path = './2019_kbo_for_kaggle_v2.csv'
```

```
In [3]: def load_data (data_path) :
            data = pd.read_csv(data_path)
            return data
```

```
In [4]: data = load_data(data_path)
```

```
In [5]: data
```

Out [5]:

	batter_name	age	G	PA	AB	R	H	2B	3B	HR	...	tp	1B	FBP	avg	OBP	SLG	OPS	p_year	YAB	YOPS
0	백용환	24.0	26.0	58.0	52.0	4.0	9.0	4.0	0.0	0.0	...	포수	5.0	6.0	0.173	0.259	0.250	0.509	2014	79.0	0.580
1	백용환	25.0	47.0	86.0	79.0	8.0	14.0	2.0	0.0	4.0	...	포수	8.0	5.0	0.177	0.226	0.354	0.580	2015	154.0	0.784
2	백용환	26.0	65.0	177.0	154.0	22.0	36.0	6.0	0.0	10.0	...	포수	20.0	20.0	0.234	0.316	0.468	0.784	2016	174.0	0.581
3	백용환	27.0	80.0	199.0	174.0	12.0	34.0	7.0	0.0	4.0	...	포수	23.0	20.0	0.195	0.276	0.305	0.581	2017	17.0	0.476
4	백용환	28.0	15.0	20.0	17.0	2.0	3.0	0.0	0.0	0.0	...	포수	3.0	3.0	0.176	0.300	0.176	0.476	2018	47.0	0.691
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1908	이원석	32.0	128.0	543.0	479.0	74.0	144.0	30.0	1.0	20.0	...	3루수	93.0	59.0	0.301	0.374	0.493	0.867	2019	395.0	0.768
1909	조용호	28.0	68.0	225.0	191.0	34.0	52.0	7.0	1.0	0.0	...	우익수	44.0	28.0	0.272	0.365	0.319	0.684	2018	13.0	0.154
1910	조용호	29.0	16.0	14.0	13.0	4.0	1.0	0.0	0.0	0.0	...	우익수	1.0	0.0	0.077	0.077	0.077	0.154	2019	188.0	0.720
1911	히메네스	27.0	70.0	299.0	279.0	37.0	87.0	17.0	2.0	11.0	...	3루수	57.0	16.0	0.312	0.344	0.505	0.849	2016	523.0	0.889
1912	히메네스	28.0	135.0	579.0	523.0	101.0	161.0	36.0	0.0	26.0	...	3루수	99.0	49.0	0.308	0.363	0.526	0.889	2017	181.0	0.769

1913 rows × 37 columns

데이터 확인


```
In [7]: print(data.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1913 entries, 0 to 1912
Data columns (total 37 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   batter_name  1913 non-null   object
 1   age          1913 non-null   float64
 2   G            1913 non-null   float64
 3   PA           1913 non-null   float64
 4   AB           1913 non-null   float64
 5   R            1913 non-null   float64
 6   H            1913 non-null   float64
 7   2B           1913 non-null   float64
 8   3B           1913 non-null   float64
 9   HR           1913 non-null   float64
 10  TB           1913 non-null   float64
 11  RBI          1913 non-null   float64
 12  SB           1913 non-null   float64
 13  CS           1913 non-null   float64
 14  BB           1913 non-null   float64
 15  HBP          1913 non-null   float64
 16  GB           1913 non-null   float64
 17  SO           1913 non-null   float64
 18  GDP          1913 non-null   float64
 19  BU           1913 non-null   float64
 20  fly          1913 non-null   float64
 21  year         1913 non-null   int64
 22  salary       1913 non-null   int64
 23  war          1913 non-null   float64
 24  year_born    1913 non-null   object
 25  hand2        1913 non-null   object
 26  cp           1913 non-null   object
 27  tp           1913 non-null   object
 28  1B           1913 non-null   float64
 29  FBP          1913 non-null   float64
 30  avg          1899 non-null   float64
 31  OBP          1901 non-null   float64
 32  SLG          1899 non-null   float64
 33  OPS          1899 non-null   float64
 34  p_year       1913 non-null   int64
 35  YAB          1913 non-null   float64
 36  YOPS         1898 non-null   float64
dtypes: float64(29), int64(3), object(5)
memory usage: 553.1+ KB
None
```

데이터 소실
(결측값) 확인

데이터 정제

In [15]:
```
# 다음 시즌 관련 데이터 삭제하기
data = data.drop(['p_year', 'YAB', 'YOPS'], axis = 1)
data
```

Out [15]:

	batter_name	age	G	PA	AB	R	H	2B	3B	HR	...	year_born	hand2	cp	tp	1B	FBP	avg	OBP	SLG	OPS
0	백용환	24.0	26.0	58.0	52.0	4.0	9.0	4.0	0.0	0.0	...	1989-03-20	우투우타	포수	포수	5.0	6.0	0.173	0.259	0.250	0.509
1	백용환	25.0	47.0	86.0	79.0	8.0	14.0	2.0	0.0	4.0	...	1989-03-20	우투우타	포수	포수	8.0	5.0	0.177	0.226	0.354	0.580
2	백용환	26.0	65.0	177.0	154.0	22.0	36.0	6.0	0.0	10.0	...	1989-03-20	우투우타	포수	포수	20.0	20.0	0.234	0.316	0.468	0.784
3	백용환	27.0	80.0	199.0	174.0	12.0	34.0	7.0	0.0	4.0	...	1989-03-20	우투우타	포수	포수	23.0	20.0	0.195	0.276	0.305	0.581
4	백용환	28.0	15.0	20.0	17.0	2.0	3.0	0.0	0.0	0.0	...	1989-03-20	우투우타	포수	포수	3.0	3.0	0.176	0.300	0.176	0.476
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1908	이원석	32.0	128.0	543.0	479.0	74.0	144.0	30.0	1.0	20.0	...	1986-10-21	우투우타	3루수	3루수	93.0	59.0	0.301	0.374	0.493	0.867
1909	조용호	28.0	68.0	225.0	191.0	34.0	52.0	7.0	1.0	0.0	...	1989-09-09	우투좌타	우익수	우익수	44.0	28.0	0.272	0.365	0.319	0.684
1910	조용호	29.0	16.0	14.0	13.0	4.0	1.0	0.0	0.0	0.0	...	1989-09-09	우투좌타	우익수	우익수	1.0	0.0	0.077	0.077	0.077	0.154
1911	히메네스	27.0	70.0	299.0	279.0	37.0	87.0	17.0	2.0	11.0	...	1988-01-18	우투우타	3루수	3루수	57.0	16.0	0.312	0.344	0.505	0.849
1912	히메네스	28.0	135.0	579.0	523.0	101.0	161.0	36.0	0.0	26.0	...	1988-01-18	우투우타	3루수	3루수	99.0	49.0	0.308	0.363	0.526	0.889

1913 rows × 34 columns

데이터 정제

In [16]:
```
# 결측치가 있는 행 제거하기.
data = data.dropna()
data
```

Out[16]:

	batter_name	age	G	PA	AB	R	H	2B	3B	HR	...	year_born	hand2	cp	tp	1B	FBP	avg	OBP	SLG	OPS
0	백용환	24.0	26.0	58.0	52.0	4.0	9.0	4.0	0.0	0.0	...	1989-03-20	우투우타	포수	포수	5.0	6.0	0.173	0.259	0.250	0.509
1	백용환	25.0	47.0	86.0	79.0	8.0	14.0	2.0	0.0	4.0	...	1989-03-20	우투우타	포수	포수	8.0	5.0	0.177	0.226	0.354	0.580
2	백용환	26.0	65.0	177.0	154.0	22.0	36.0	6.0	0.0	10.0	...	1989-03-20	우투우타	포수	포수	20.0	20.0	0.234	0.316	0.468	0.784
3	백용환	27.0	80.0	199.0	174.0	12.0	34.0	7.0	0.0	4.0	...	1989-03-20	우투우타	포수	포수	23.0	20.0	0.195	0.276	0.305	0.581
4	백용환	28.0	15.0	20.0	17.0	2.0	3.0	0.0	0.0	0.0	...	1989-03-20	우투우타	포수	포수	3.0	3.0	0.176	0.300	0.176	0.476
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1908	이원석	32.0	128.0	543.0	479.0	74.0	144.0	30.0	1.0	20.0	...	1986-10-21	우투우타	3루수	3루수	93.0	59.0	0.301	0.374	0.493	0.867
1909	조용호	28.0	68.0	225.0	191.0	34.0	52.0	7.0	1.0	0.0	...	1989-09-09	우투좌타	우익수	우익수	44.0	28.0	0.272	0.365	0.319	0.684
1910	조용호	29.0	16.0	14.0	13.0	4.0	1.0	0.0	0.0	0.0	...	1989-09-09	우투좌타	우익수	우익수	1.0	0.0	0.077	0.077	0.077	0.154
1911	히메네스	27.0	70.0	299.0	279.0	37.0	87.0	17.0	2.0	11.0	...	1988-01-18	우투우타	3루수	3루수	57.0	16.0	0.312	0.344	0.505	0.849
1912	히메네스	28.0	135.0	579.0	523.0	101.0	161.0	36.0	0.0	26.0	...	1988-01-18	우투우타	3루수	3루수	99.0	49.0	0.308	0.363	0.526	0.889

1899 rows × 34 columns

In [17]:
```
data.isnull().values.any()
```

Out[17]:
```
False
```

데이터 열 추가

	batter_name	age	G	PA	AB	R	H	2B	3B	HR	...	1B	FBP	avg	OBP	SLG	OPS	GPA	RC	XR	wOBA
0	백용환	24.0	26.0	58.0	52.0	4.0	9.0	4.0	0.0	0.0	...	5.0	6.0	0.173	0.259	0.250	0.509	0.7162	1.980000	1.772	0.242563
1	백용환	25.0	47.0	86.0	79.0	8.0	14.0	2.0	0.0	4.0	...	8.0	5.0	0.177	0.226	0.354	0.580	0.7608	2.388387	5.906	0.253297
2	백용환	26.0	65.0	177.0	154.0	22.0	36.0	6.0	0.0	10.0	...	20.0	20.0	0.234	0.316	0.468	0.784	1.0368	8.125000	21.594	0.330590
3	백용환												20.0	0.195	0.276	0.305	0.581	0.8018	7.759638	12.084	0.257986
4	백용환												3.0	0.176	0.300	0.176	0.476	0.7160	0.657391	0.226	0.246261
...	...												...	...	...	...	...	...	...	...	...
1908	이원석												59.0	0.301	0.374	0.493	0.867	1.1662	43.000840	76.986	0.364215
1909	조용호												28.0	0.272	0.365	0.319	0.684	0.9760	18.676328	20.798	0.303364
1910	조용호												0.0	0.077	0.077	0.077	0.154	0.2156	0.000000	-0.990	0.063429
1911	히메네스												16.0	0.312	0.344	0.505	0.849	1.1242	24.825916	43.986	0.354604
1912	히메네스												49.0	0.308	0.363	0.526	0.889	1.1794	45.676090	87.444	0.368883

타자에 관한 세이버메트릭스 지표들

- OPS = 출루율 + 장타율
- RC = [(안타 + 볼넷 - 도실 - 병살타)x (루타수+0.52x(도루+희생타)+ 0.26x(볼넷 - 고의볼넷))] / (타수+볼넷+희생타)
- XR = (1루타×0.5) + (2루타×0.72) + (3루타×1.04) + (홈런×1.44) + ((4사구-고의4구)×0.34) + (고의4구×0.25) + (도루×0.18) - (도실×0.32) - ((타수 - 안타 - 삼진)×0.09) - (삼진×0.098) - (병살타×0.37) + (희비×0.37) + (희타×0.04)
- wOBA = [((볼넷 - 고의4구)×0.69) + (몸에 맞은 공×0.722) + (1루타 ×0.888) + (2루타×1.271) + (3루타×1.616) + (홈런×2.101)] /(타수 + 볼넷 - 고의4구 + 희비 + 몸에 맞은 공)
- WAR = (공격지표 + 주루지표 + 수비지표 + 포지션 보정 + 대체선수 대비 타석수 보정)/(승리당 득점)

새로운 데이터셋 생성

In [27]:
```
# 필요한 열만 추출하기
data_column = ['batter_name', 'age', 'G', 'AB', 'RBI', 'year', 'salary', 'war', 'hand2',
               'tp', 'avg', 'OBP', 'SLG', 'OPS', 'GPA', 'RC', 'XR', 'wOBA']

final_data = data[data_column]
final_data
```

Out [27]:

	batter_name	age	G	AB	RBI	year	salary	war	hand2	tp	avg	OBP	SLG	OPS	GPA	RC	XR	wOBA
0	백용환	24.0	26.0	52.0	3.0	2013	2500	-0.055	우투우타	포수	0.173	0.259	0.250	0.509	0.7162	1.980000	1.772	0.242563
1	백용환	25.0	47.0	79.0	10.0	2014	2900	-0.441	우투우타	포수	0.177	0.226	0.354	0.580	0.7608	2.388387	5.906	0.253297
2	백용환	26.0	65.0	154.0	30.0	2015	6000	0.783	우투우타	포수	0.234	0.316	0.468	0.784	1.0368	8.125000	21.594	0.330590
3	백용환	27.0	80.0	174.0	15.0	2016	6000	-0.405	우투우타	포수	0.195	0.276	0.305	0.581	0.8018	7.759638	12.084	0.257986
4	백용환	28.0	15.0	17.0	1.0	2017	5500	-0.130	우투우타	포수	0.176	0.300	0.176	0.476	0.7160	0.657391	0.226	0.246261
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1908	이원석	32.0	128.0	479.0	93.0	2018	30000	3.315	우투우타	3루수	0.301	0.374	0.493	0.867	1.1662	43.000840	76.986	0.364215
1909	조용호	28.0	68.0	191.0	10.0	2017	3100	0.215	우투좌타	우익수	0.272	0.365	0.319	0.684	0.9760	18.676328	20.798	0.303364
1910	조용호	29.0	16.0	13.0	0.0	2018	6200	-0.271	우투좌타	우익수	0.077	0.077	0.077	0.154	0.2156	0.000000	-0.990	0.063429
1911	히메네스	27.0	70.0	279.0	46.0	2015	50000	2.365	우투우타	3루수	0.312	0.344	0.505	0.849	1.1242	24.825916	43.986	0.354604
1912	히메네스	28.0	135.0	523.0	102.0	2016	50000	5.356	우투우타	3루수	0.308	0.363	0.526	0.889	1.1794	45.676090	87.444	0.368883

1899 rows × 18 columns

새로운 데이터셋 확인


```
In [29]: final_data['batter_name'].nunique()
Out[29]: 336

In [30]: name_length = final_data.groupby('batter_name').size()
         name_length
Out[30]: batter_name
         강경학      5
         강동관      1
         강동우     13
         강민국      3
         강민호     15
                 ..
         황선일      3
         황윤호      4
         황재균     10
         황진수      6
         히메네스     2
         Length: 336, dtype: int64
```

총 336명의 선수들이 존재하는 것을 확인할 수 있고 각자의 데이터수가 생각보다 고르지 않는 것을 볼 수 있다.

선수 기량 추이 분석은 2개 이하 데이터셋은 불필요

```
In [31]: name_length.median(), name_length.mean()
Out[31]: (4.0, 5.651785714285714)

In [32]: name_length.min(), name_length.max()
Out[32]: (1, 19)
```

4.
데이터 분석(1)

야구선수의 기량 성장 추이 분석

추가 전처리

1.1 필요없는 데이터 삭제

In [33]:
```
# 출전횟수가 2회 이하인 지표는 제외하기
data1 = final_data.groupby('batter_name').filter(lambda x: x['age'].count() > 2)
data1
```

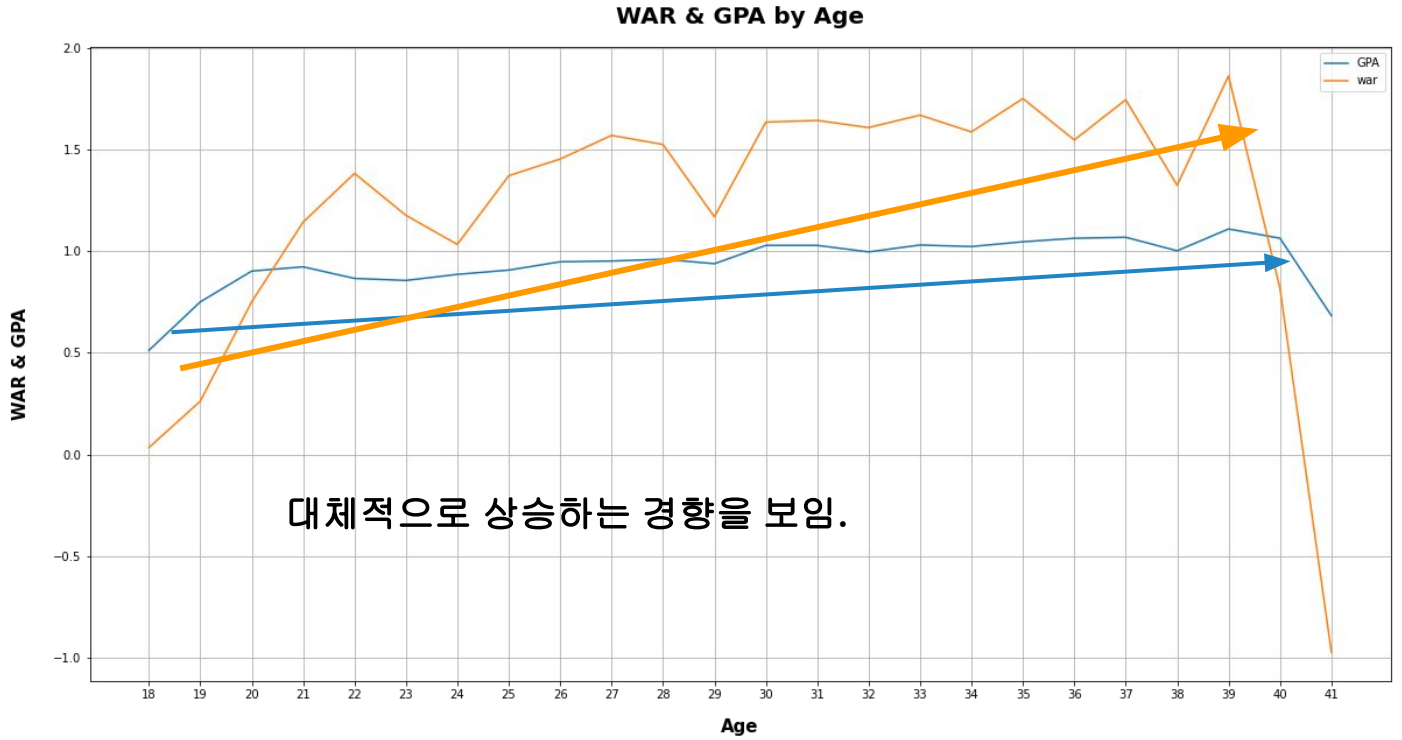
Out [33]:

	batter_name	age	G	AB	RBI	year	salary	war	hand2	tp	avg	OBP	SLG	OPS	GPA	RC	XR	wOBA
0	백용환	24.0	26.0	52.0	3.0	2013	2500	-0.055	우투우타	포수	0.173	0.259	0.250	0.509	0.7162	1.980000	1.772	0.242563
1	백용환	25.0	47.0	79.0	10.0	2014	2900	-0.441	우투우타	포수	0.177	0.226	0.354	0.580	0.7608	2.388387	5.906	0.253297
2	백용환	26.0	65.0	154.0	30.0	2015	6000	0.783	우투우타	포수	0.234	0.316	0.468	0.784	1.0368	8.125000	21.594	0.330590
3	백용환	27.0	80.0	174.0	15.0	2016	6000	-0.405	우투우타	포수	0.195	0.276	0.305	0.581	0.8018	7.759638	12.084	0.257986
4	백용환	28.0	15.0	17.0	1.0	2017	5500	-0.130	우투우타	포수	0.176	0.300	0.176	0.476	0.7160	0.657391	0.226	0.246261
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1904	이원석	26.0	107.0	325.0	42.0	2012	8400	2.056	우투우타	3루수	0.268	0.337	0.394	0.731	1.0006	22.203883	33.950	0.306778
1905	이원석	27.0	85.0	264.0	39.0	2013	10000	2.810	우투우타	3루수	0.314	0.385	0.473	0.858	1.1660	25.764182	41.838	0.362973
1906	이원석	30.0	7.0	19.0	7.0	2016	15000	0.311	우투우타	3루수	0.316	0.364	0.789	1.153	1.4442	1.318333	4.764	0.436120
1907	이원석	31.0	121.0	411.0	62.0	2017	30000	1.102	우투우타	3루수	0.265	0.323	0.450	0.773	1.0314	28.650221	56.860	0.324395
1908	이원석	32.0	128.0	479.0	93.0	2018	30000	3.315	우투우타	3루수	0.301	0.374	0.493	0.867	1.1662	43.000840	76.986	0.364215

1738 rows × 18 columns

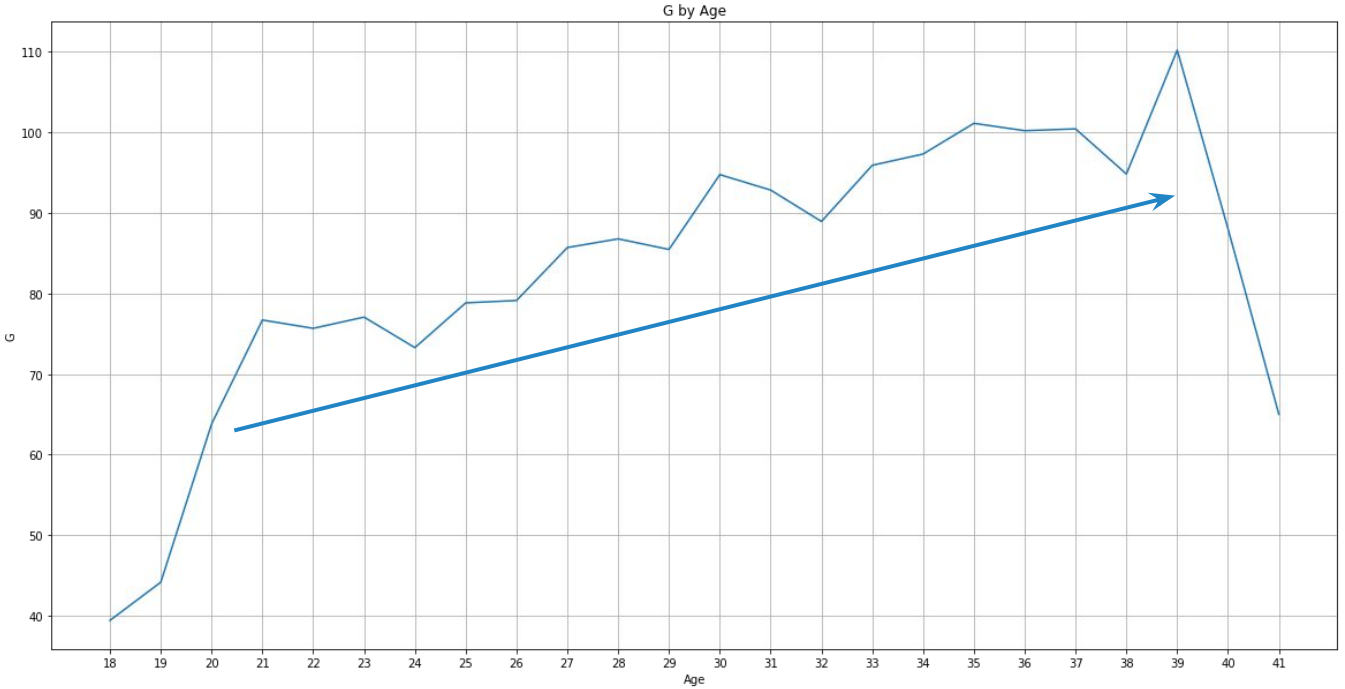
연령별 기량 추이 그래프

그래프 그리기



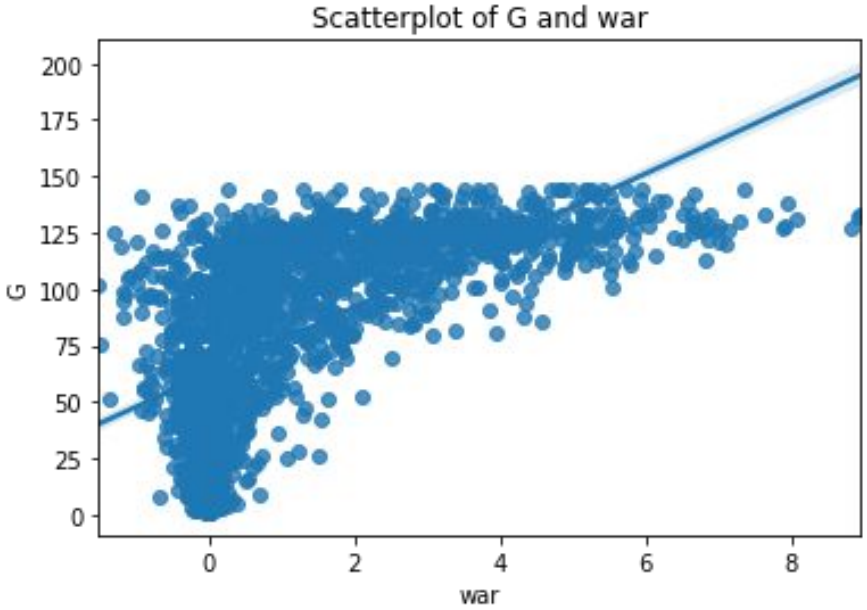
연령별 경기수 추이 그래프

그래프 그리기



연령별 경기수 추이 그래프

그래프 그리기



출전 횟수와 선수의 기량은 매우 강력한 양의 상관관계가 있다.

=> 타자는 노화보단 경험치가 더 크게 작용한다.

타자는 경험치에 따라 기량 UP!

40살 즈음에 전성기를 맞이하게 된다.

5.
데이터 분석(2)

야구선수의 기량에 따른 연봉 적합도 분석

추가 전처리

2.1 데이터 추가 전처리

우선 각 연도별 선수 데이터 길이를 확인한다.

```
]: final_data[['year' , 'batter_name']].groupby('year').nunique()
```

year	batter_name
1990	8
1991	6
1992	4
1993	4
1994	5
1995	8
1996	12
1997	15
1998	19
1999	21
2000	25
2001	32
2002	35
2003	43
2004	51
2005	54
2006	65
2007	67
2008	79
2009	87
2010	88
2011	100
2012	119
2013	122
2014	133
2015	160
2016	169
2017	174
2018	194

```
In [43]: year_length = data.groupby('year').size()
         year_length.median(), year_length.mean()
Out[43]: (51.0, 65.48275862068965)
```

- 타자의 전체 수는 약 300명
- 그에 비해 너무 적은 데이터셋은 분석에 큰 도움이 되기 어려움.
- 따라서 최근 10년간의 데이터만 잡아 분석을 진행함.

추가 전처리

In [44]:
```
# 최근 10년간의 데이터만 추출하기
data2 = final_data[final_data['year']> 2008]
data2
```

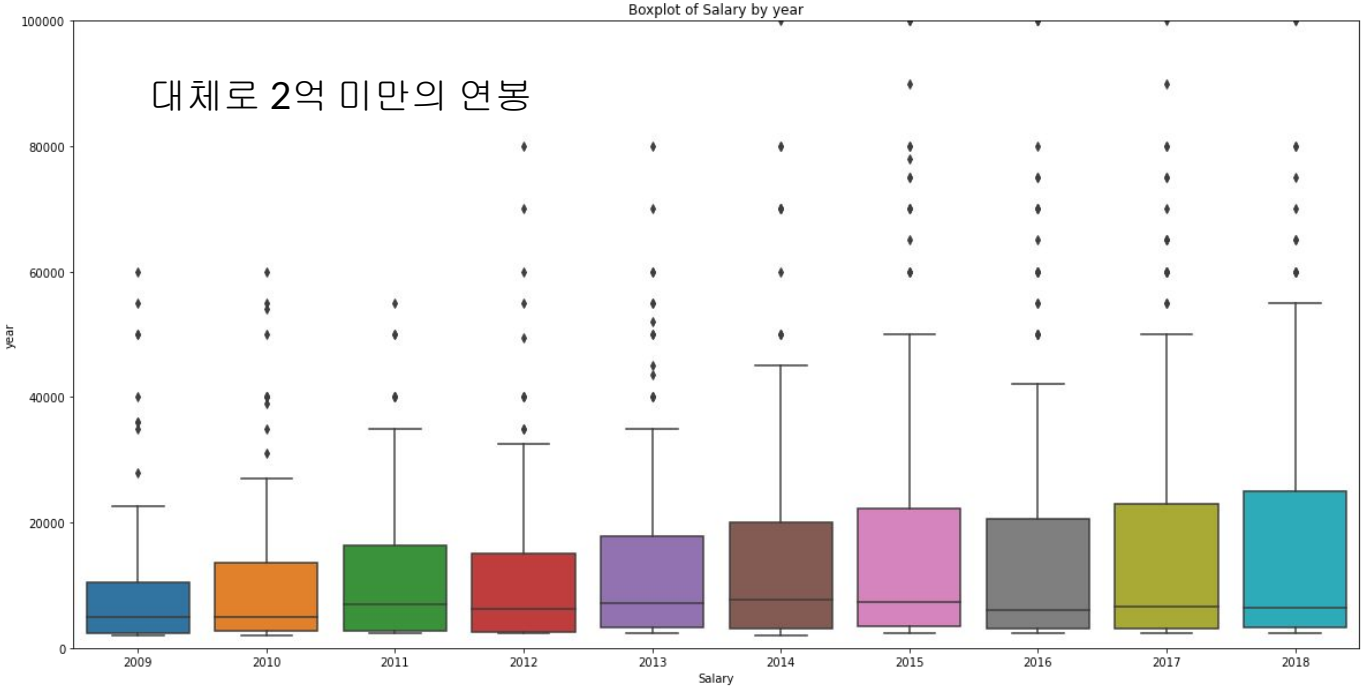
Out [44]:

	batter_name	age	G	AB	RBI	year	salary	war	hand2	tp	avg	OBP	SLG	OPS	GPA	RC	XR	wOBA
0	백용환	24.0	26.0	52.0	3.0	2013	2500	-0.055	우투우타	포수	0.173	0.259	0.250	0.509	0.7162	1.980000	1.772	0.242563
1	백용환	25.0	47.0	79.0	10.0	2014	2900	-0.441	우투우타	포수	0.177	0.226	0.354	0.580	0.7608	2.388387	5.906	0.253297
2	백용환	26.0	65.0	154.0	30.0	2015	6000	0.783	우투우타	포수	0.234	0.316	0.468	0.784	1.0368	8.125000	21.594	0.330590
3	백용환	27.0	80.0	174.0	15.0	2016	6000	-0.405	우투우타	포수	0.195	0.276	0.305	0.581	0.8018	7.759638	12.084	0.257986
4	백용환	28.0	15.0	17.0	1.0	2017	5500	-0.130	우투우타	포수	0.176	0.300	0.176	0.476	0.7160	0.657391	0.226	0.246261
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1908	이원석	32.0	128.0	479.0	93.0	2018	30000	3.315	우투우타	3루수	0.301	0.374	0.493	0.867	1.1662	43.000840	76.986	0.364215
1909	조용호	28.0	68.0	191.0	10.0	2017	3100	0.215	우투좌타	우익수	0.272	0.365	0.319	0.684	0.9760	18.676328	20.798	0.303364
1910	조용호	29.0	16.0	13.0	0.0	2018	6200	-0.271	우투좌타	우익수	0.077	0.077	0.077	0.154	0.2156	0.000000	-0.990	0.063429
1911	히메네스	27.0	70.0	279.0	46.0	2015	50000	2.365	우투우타	3루수	0.312	0.344	0.505	0.849	1.1242	24.825916	43.986	0.354604
1912	히메네스	28.0	135.0	523.0	102.0	2016	50000	5.356	우투우타	3루수	0.308	0.363	0.526	0.889	1.1794	45.676090	87.444	0.368883

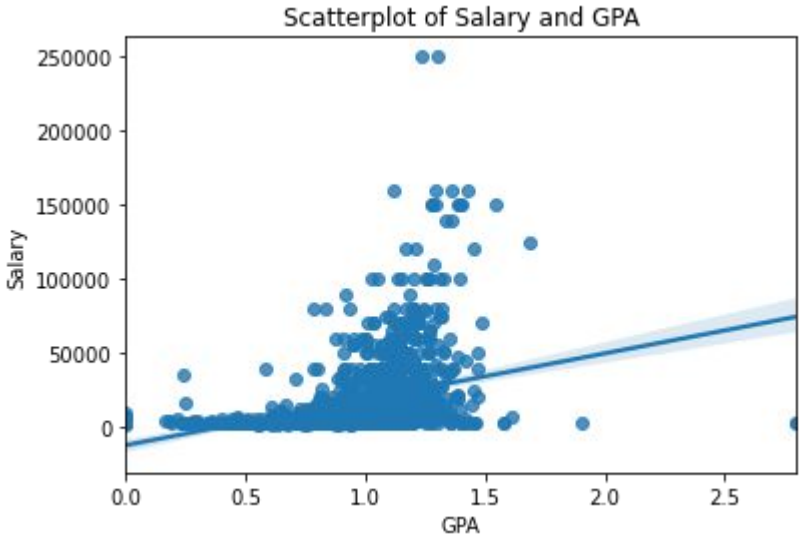
1346 rows × 18 columns

연봉 기준점 정하기

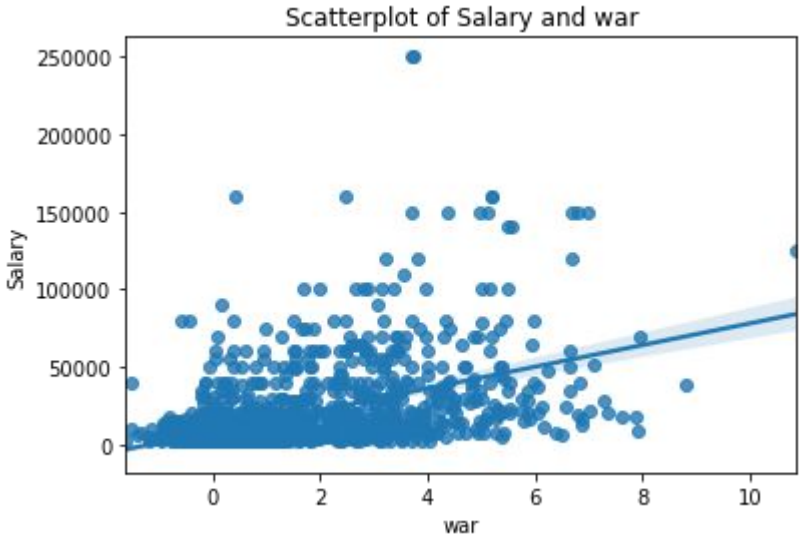
데이터 확인



연봉과 기량의 관계




```
In [48]: stats.pearsonr(data['GPA'], data['salary'])
Out[48]: (0.3205878060535939, 1.205497151706147e-46)
```



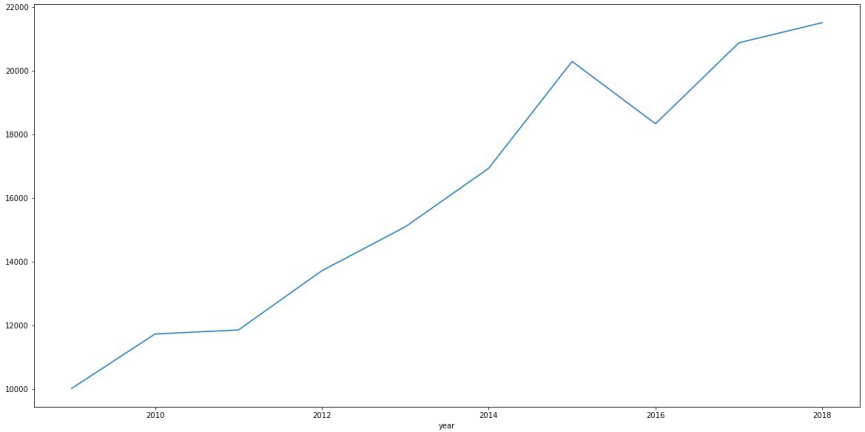

```
In [50]: stats.pearsonr(data['war'], data['salary'])
Out[50]: (0.45073191105665883, 1.1453467742138795e-95)
```

= 강한 양의 상관관계가 있다.

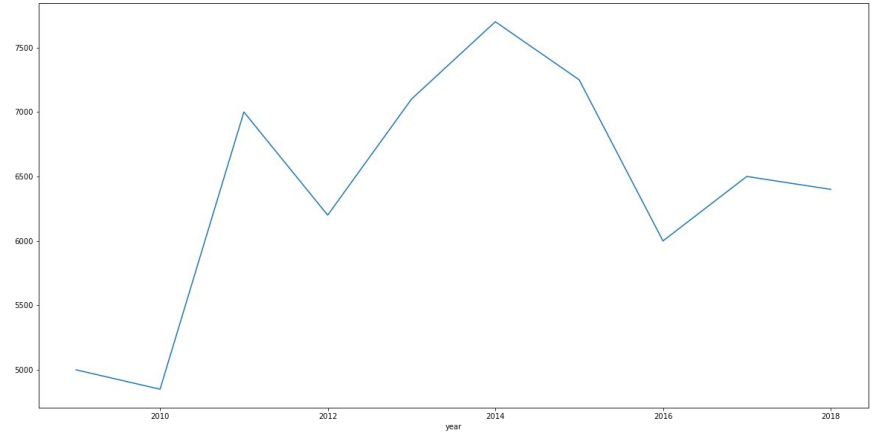
연봉과 기량의 대표 지표

데이터 확인

연도별 연봉 평균



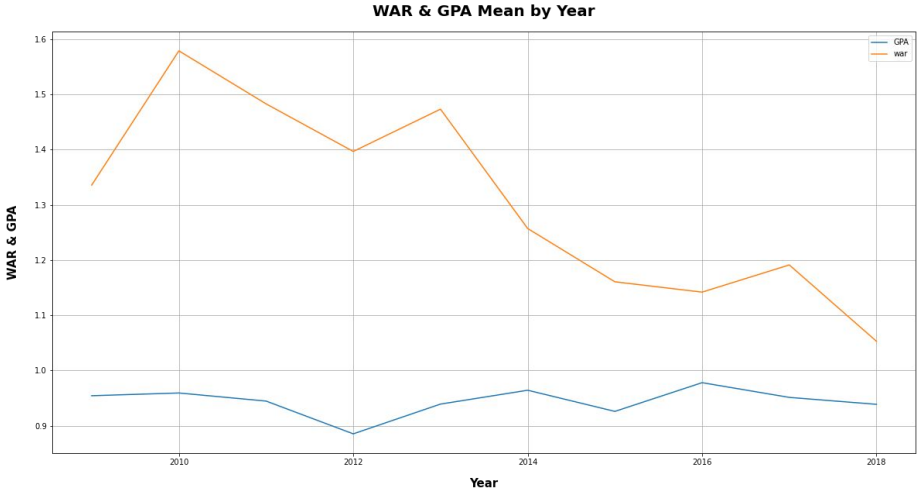
연도별 연봉 중간값



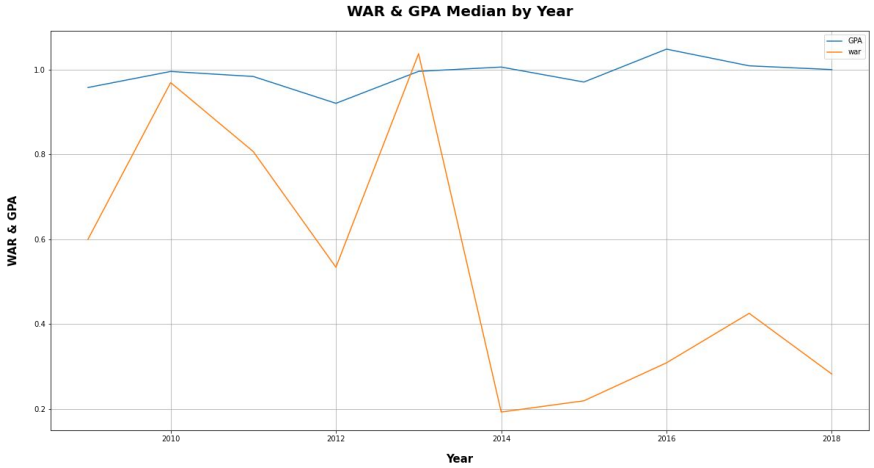
연봉과 기량의 대표 지표

데이터 확인

연도별 기량 평균



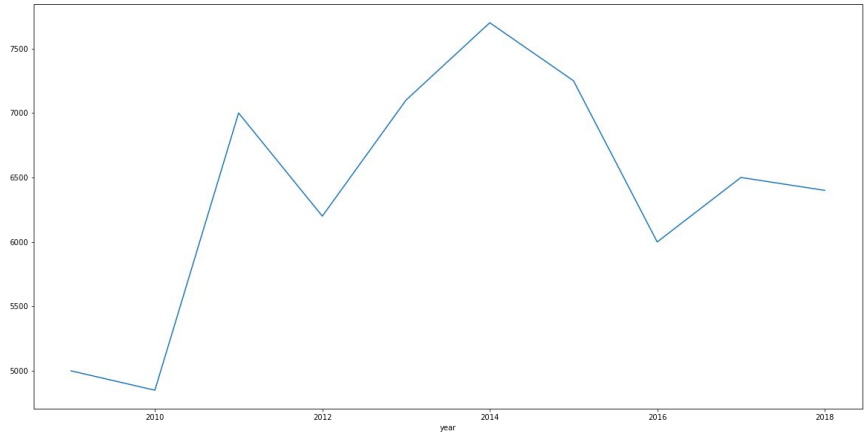
연도별 기량 중간값



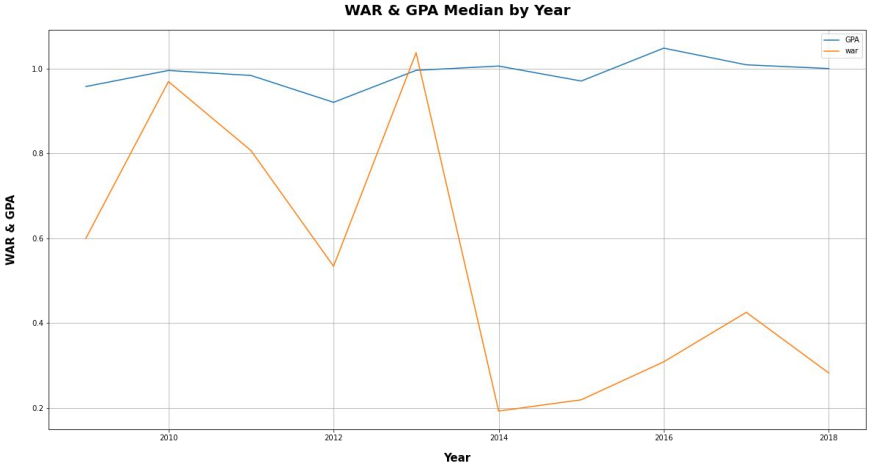
연봉과 기량의 대표 지표

데이터 확인

연도별 연봉 중간값



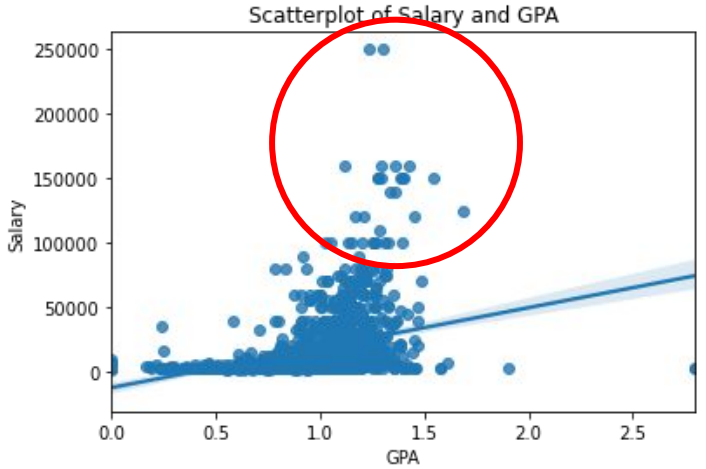
연도별 기량 중간값



데이터의 추이가 **GPA와 연봉의 중간값**이 매우 유사함.

이런 결과가 나온 이유?

- GPA의 평균 산점도는 예외값이 상당히 많기 때문
- 이로 인해 평균의 함정이 생김
- 연봉은 WAR보단 GPA를 더 잘 따라가는 것으로 보임



과대평가 된 사람 확인

필터링

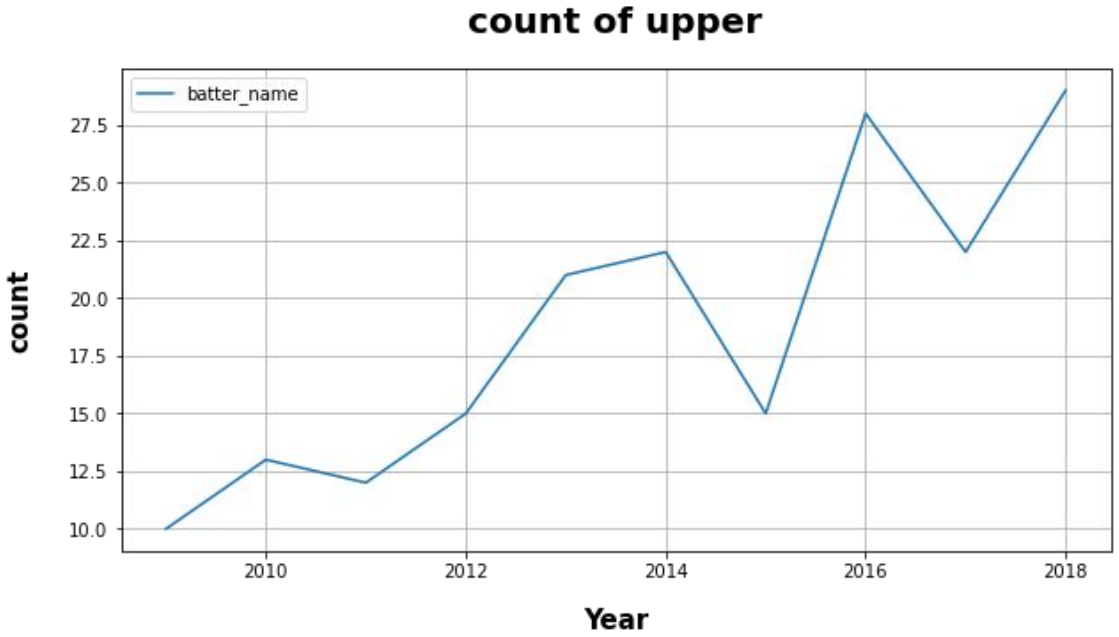
In [65]:
```
# 과평가된 사람들 필터링하기
for i in range(2009, 2019) :
    data4 = data3[data3['year'] == i]
    data5 = data4[data4['GPA'] < y_gpa_median[i]]
    data6 = data5[data5['salary'] > salary_median[i]]
    if i == 2009 :
        data7 = data6
    else :
        data7 = data7.append(data6, ignore_index=True)
```

In [66]:
```
data7
```

Out [66]:

	batter_name	age	RBI	year	salary	tp	avg	GPA	war
0	이현곤	29.0	33.0	2009	10000	3루수	0.253	0.9032	0.600
1	조인성	34.0	36.0	2009	40000	포수	0.214	0.9390	0.889
2	조동화	28.0	8.0	2009	10500	좌익수	0.178	0.6474	-1.523
3	안치용	30.0	30.0	2009	6600	좌익수	0.237	0.9110	-0.142
4	진갑용	35.0	20.0	2009	50000	포수	0.232	0.9088	0.381
...	...	...	...	...	...	...	...	...	...
182	김상수	28.0	50.0	2018	24000	유격수	0.263	0.9272	1.039
183	김성현	31.0	55.0	2018	25000	유격수	0.277	0.9618	0.698
184	이상호	29.0	8.0	2018	7500	3루수	0.251	0.7822	-0.758
185	이해창	31.0	24.0	2018	7100	포수	0.216	0.8678	-0.134
186	최원준	21.0	32.0	2018	7500	우익수	0.272	0.9472	-0.099

187 rows × 9 columns



과소평가 된 사람 확인

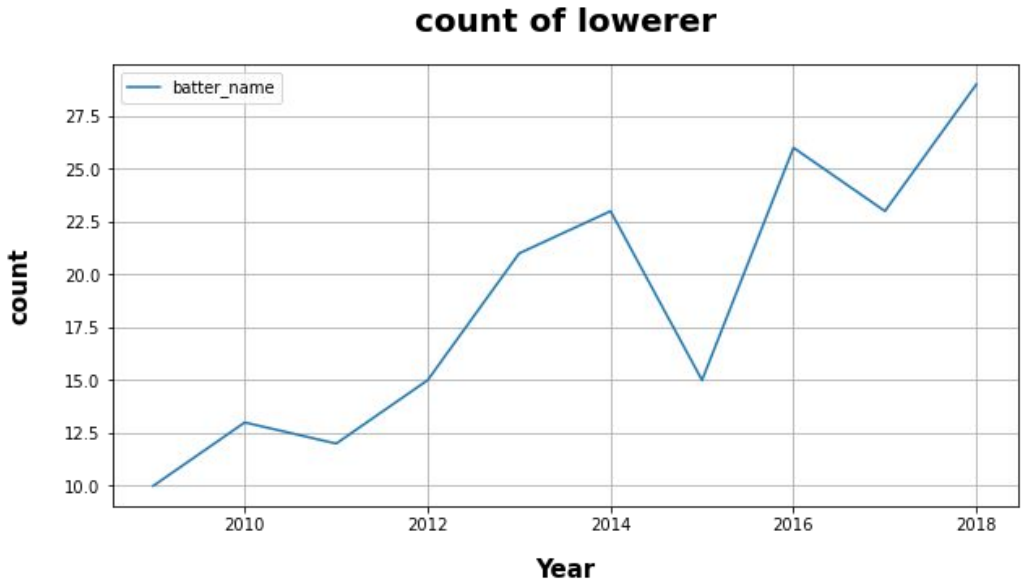
In [69]:
```
# 과소평가된 사람들 필터링하기
data8 = data6
for i in range(2009, 2019) :
    data4 = data3[data3['year'] == i]
    data5 = data4[data4['GPA'] > y_gpa_median[i]]
    data6 = data5[data5['salary'] < salary_median[i]]
    if i == 2009 :
        data8 = data6
    else :
        data8 = data8.append(data6, ignore_index=True)
```

In [70]:
```
data8
```

Out [70]:

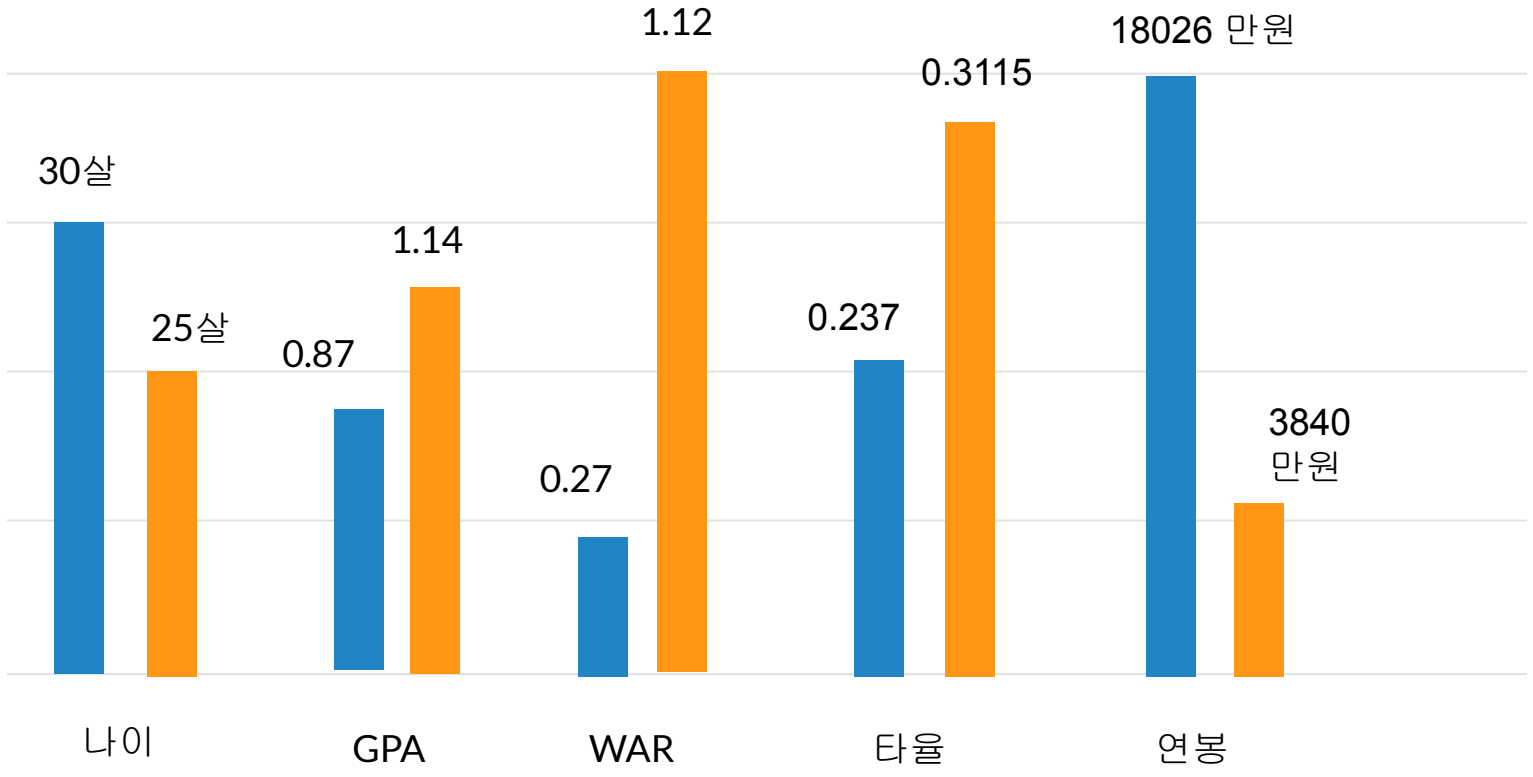
	batter_name	age	RBI	year	salary	tp	avg	GPA	war
0	김선빈	20.0	6.0	2009	3500	유격수	0.293	1.0252	0.777
1	나지완	24.0	73.0	2009	3800	지명타자	0.263	1.1364	2.360
2	정수빈	19.0	17.0	2009	2400	중견수	0.264	1.0020	1.109
3	윤진호	23.0	2.0	2009	2400	2루수	0.333	1.2722	0.057
4	김민성	21.0	37.0	2009	2300	3루수	0.248	0.9642	0.857
...	...	...	...	...	...	...	...	...	...
182	전병우	26.0	13.0	2018	2700	3루수	0.364	1.4016	1.117
183	나경민	27.0	1.0	2018	5800	중견수	0.263	1.0180	0.349
184	허일	26.0	0.0	2018	2700	중견수	0.357	1.0770	-0.016
185	박준태	27.0	24.0	2018	3800	우익수	0.228	1.0280	0.070
186	이원재	29.0	19.0	2018	2700	좌익수	0.304	1.1106	0.350

187 rows × 9 columns



두 집단 비교

데이터 분석


30살
25살
0.87
1.14
0.27
1.12
0.237
0.3115
18026 만원
3840
만원
나이
GPA
WAR
타율
연봉

젊은 세대의 역량보단 원로 세대의
네임밸류를 우대하는 등 연봉에 크게
작용하는 양극화 현상을 확인

6.

결론

분석 정리 및 기대효과

기대효과

- 의외로 야구는 나이보단 경험치가 기량이 더 크게 작용하는 것을 확인
- 야구가 연봉을 선정함에 있어서는 물론 전반적으로는 기량에 비례하게 책정
- 기량이 없음에도 **네임밸류와 경력 때문에 높은 연봉을 받는**이가 있는 사람이 있는가 하면, 반대로 **젊으면서 기량이 출중**하고, 기존 원로 멤버를 위협할 정도가 됨에도 불구하고 **연봉을 적게 받는 양극화**가 있는 것을 확인

이번 자료를 바탕으로 **과소평가된 선수는 더 좋은 연봉 협상**을 노릴 수 있고, 구단은 과대평가된 선수의 합당한 연봉을 제시할 수 있게됨.

Thanks!