22시즌 KBO 데이터 분석



2020312163 김지윤



안내 목차

연구문제 & 가설 분석 데이터 & 기법

 연구결과

 활용방안

연구 한계 & 후속 연구 제안

연구문제 & 가설

- 1. Best 10 Player 및 구단별 주요 실적 분석
- 2. 야구 선수들은 어떤 계절에 강해질까?
- 3. 득점과 여러 요소들의 상관관계



1. Best 10 Player 및 구단별 주요 실적 분석

22년도 KBO 정규 시즌(4/2~10/31) 10개 구단 60타석 이상의 선수들의 윌별 데이터 (출처 : 스탯티즈)

이름	팀	월	타석	타수	득	점 인	Eł	2Eł	3Eł	홈런	루타	타점	도루	도실	볼넷	사구	고4	삼진	병살	희타	희비	타율	출	루율	장타율 (OPS
문성주	22L		4	75	60	10	2	26	7	1	0	35	6	3	2	11	0	0	9	3	3	1 0	.433	0.514	0.583	1.097
한동희	22롯		4	101	89	16	3	38	9	0	7	68	22	0	0	10	1	1	8	3	0	1 0	.427	0.485	0.764	1.249
한유섬	225		4	103	86	17	3	34	13	1	3	58	27	0	0	12	4	2	17	0	0	1 0	.395	0.485	0.674	1.16
피렐라	22삼		4	109	100	17	3	39	8	1	2	55	15	3	0	6	3	1	13	2	0	0	0.39	0.44	0.55	0.99
이대호	22롯		4	100	90	13	3	32	2	0	2	40	10	0	0	7	1	2	9	4	0	2 0	.356	0.4	0.444	0.844
박건우	22N		4	107	94	11	3	33	5	0	1	41	16	0	3	9	2	0	13	2	0	2 0	.351	0.411	0.436	0.847
최정	22S		4	78	67	12	2	23	4	0	2	33	11	1	0	8	2	0	13	0	0	1 0	.343	0.423	0.493	0.916
류지혁	22기		4	66	62	8	2	21	2	0	1	26	10	0	0	4	0	0	13	0	0	0 0	.339	0.379	0.419	0.798
김선빈	22기		4	99	87	11	2	29	7	0	1	39	10	2	1	11	0	0	7	4	1	0 0	.333	0.408	0.448	0.856
나성범	22기		4	106	91	11	3	30	9	1	2	47	11	0	0	12	3	0	23	2	0	0	0.33	0.425	0.516	0.941
심우준	22K		4	75	61	11	2	20	0	0	0	20	5	5	1	10	1	0	11	0	3	0 0	.328	0.431	0.328	0.758
전준우	22롯		4	84	77	12	2	25	3	0	1	31	11	1	2	4	1	1	9	3	0	2 0	.325	0.357	0.403	0.76
이정후	227		4	106	96	11	3	31	7	0	4	50	20	1	0	8	1	2	3	2	0	1 0	.323	0.377	0.521	0.898
김인태	22두		4	107	90	12	2	29	2	0	1	34	12	1	0	15	2	1	13	1	0	0 0	.322	0.43	0.378	0.808
허경민	22두		4	91	80	9	2	25	6	0	0	31	7	2	0	7	2	0	7	1	1	1 0	.313	0.378	0.388	0.765
안치홍	22롯		4	100	94	15	2	29	8	2	1	44	8	1	1	5	0	0	8	2	0	1 0	.309	0.34	0.468	0.808
터크먼	22한		4	108	98	12	3	30	7	0	1	40	4	8	0	9	0	0	20	2	0	1 0	.306	0.361	0.408	0.769
노시환	22한		4	102	85	12	2	26	5	0	2	37	16	1	2	16	0	0	20	1	0	1 0	.306	0.412	0.435	0.847
홍창기	22L		4	84	72	12	2	22	4	1	0	28	7	2	1	7	3	0	8	3	1	1 0	.306	0.386	0.389	0.774
최지훈	22S		4	111	97	18	2	29	6	0	1	38	7	5	0	9	4	0	20	2	1	0 0	.299	0.382	0.392	0.774
김현수	22L		4	105	91	16	2	27	5	0	5	47	16	0	1	11	3	3	12	1	0	0 0	.297	0.391	0.516	0.907
페르난데	스22두		4	103	95	6	2	28	5	0	0	33	10	0	0	7	0	0	12	9	0	1 0	.295	0.34	0.347	0.687
박성한	22S		4	93	85	10	2	25	3	2	2	38	12	1	1	7	0	0	19	0	1	0 0	.294	0.348	0.447	0.795
오윤석	22K		4	75	65	8	1	19	5	0	1	27	11	1	1	7	2	0	20	2	0	1 0	.292	0.373	0.415	0.789
손아섭	22N		4	113	100	10	2	29	6	0	0	35	5	1	1	10	1	1	17	0	0	2	0.29	0.354	0.35	0.704
황재균	22K		4	111	98	11	2	28	5	1	2	41	11	3	1	11	1	1	18	1	1	0 0	.286	0.364	0.418	0.782
김혜성	227		4	112	107	13	3	80	3	1	0	35	10	8	1	5	0	0	22	3	0	0	0.28	0.313	0.327	0.64
유강남	22L		4	94	86	7	2	24	3	0	1	30	12	0	0	6	1	0	24	1	1	0 0	.279	0.333	0.349	0.682
김강민	22S		4	62	54	5	1	15	2	0	0	17	4	1	2	6	1	0	10	1	0	1 0	.278	0.355	0.315	0.67
김민혁	22K		4	91	83	12	2	23	3	0	0	26	8	2	1	7	0	0	13	1	0	1 0	.277	0.33	0.313	0.643
박동원	22기		4	60	51	5	1	14	5	0	2	25	7	0	0	9	0	1	12	0	0	0 0	.275	0.383	0.49	0.874
문보경	22L		4	75	69	8	1	18	1	0	2	25	10	0	0	4	1	0	14	2	0	1 0	.261	0.307	0.362	0.669
황대인	22기		4	99	89	4	2	23	4	0	1	30	13	0	0	5	4	0	18	4	0	1 0	.258	0.323	0.337	0.66
크론	225		4	104	98	8	2	25	6	0	4	43	17	0	0	3	1	0	24	0	0	2 0	.255	0.279	0.439	0.718

"세이버메트릭스 기법"

야구를 통계학적 · 수학적 방법으로 분석하는 방법론

1. Best 10 Player 및 구단별 주요 실적 분석

	팀	이름	루타	볼넷	사구	안타	타수	타점	홈런	희비
0	KIA	김도영	20	2	1	15	84	4	0	0
1	KIA	김석환	13	7	2	9	52	3	1	0
2	KIA	김선빈	177	65	6	145	505	61	3	3
3	KIA	나성범	286	64	17	180	563	97	21	5
4	KIA	류지혁	141	56	6	112	405	48	2	2
117	한화	장진혁	8	1	0	8	55	3	0	1
118	한화	정은원	188	85	1	140	508	49	8	3
119	한화	최재훈	110	44	21	81	364	30	5	2
120	한화	터크먼	247	64	7	166	575	43	12	2
121	한화	하주석	142	25	3	109	400	53	5	2

1. Best 10 Player 및 구단별 주요 실적 분석

```
타율 = 안타 : 타수
출루율 = (안타+볼넷+사구) : (타수+볼넷+사구+희비)
장타율 = (1루타+2×2루타+3×3루타+4×홈런) : 타수
OPS = 출루율 + 장타율
```

```
#단율, 출루율, 장단율, OPS를 계산해주는 함수 생성

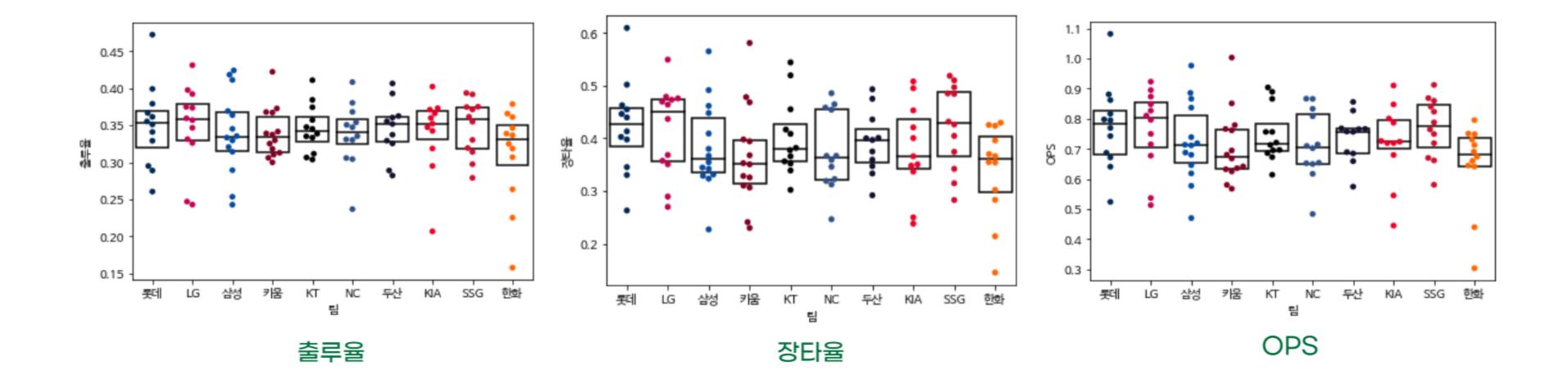
def cal_hit(df):
    df['타윤'] = df['안타'] / df['타우']
    df['출루윤'] = (df['안타'] + df['볼뎃'] + df['사구']) / (df['타우'] + df['볼뎃'] + df['사구'] + df['희비'])
    df['장타윤'] = df['루타'] / df['타우']
    df['OPS'] = df['출루윤'] + df['장타윤']
    return df

player_stat = cal_hit(data_player)
player_stat
#출루율, 장타율, OPS, 타율 순서대로 best 10 players

player_stat = player_stat.sort_values(by = ['출루윤', '장타윤', 'OPS', '타윤'], ascending = False)
player_stat = player_stat.reset_index(drop = True)
player_stat.head(10)
player_stat['이름'][:10]
```

```
0 고승민
1 문성주
2 강한울
3 이정후
4 김재성
5 괴렐라
5 김준태
7 박건우
8 김인태
9 나성범
Name: 이름, dtype: object
```

1. Best 10 Player 및 구단별 주요 실적 분석

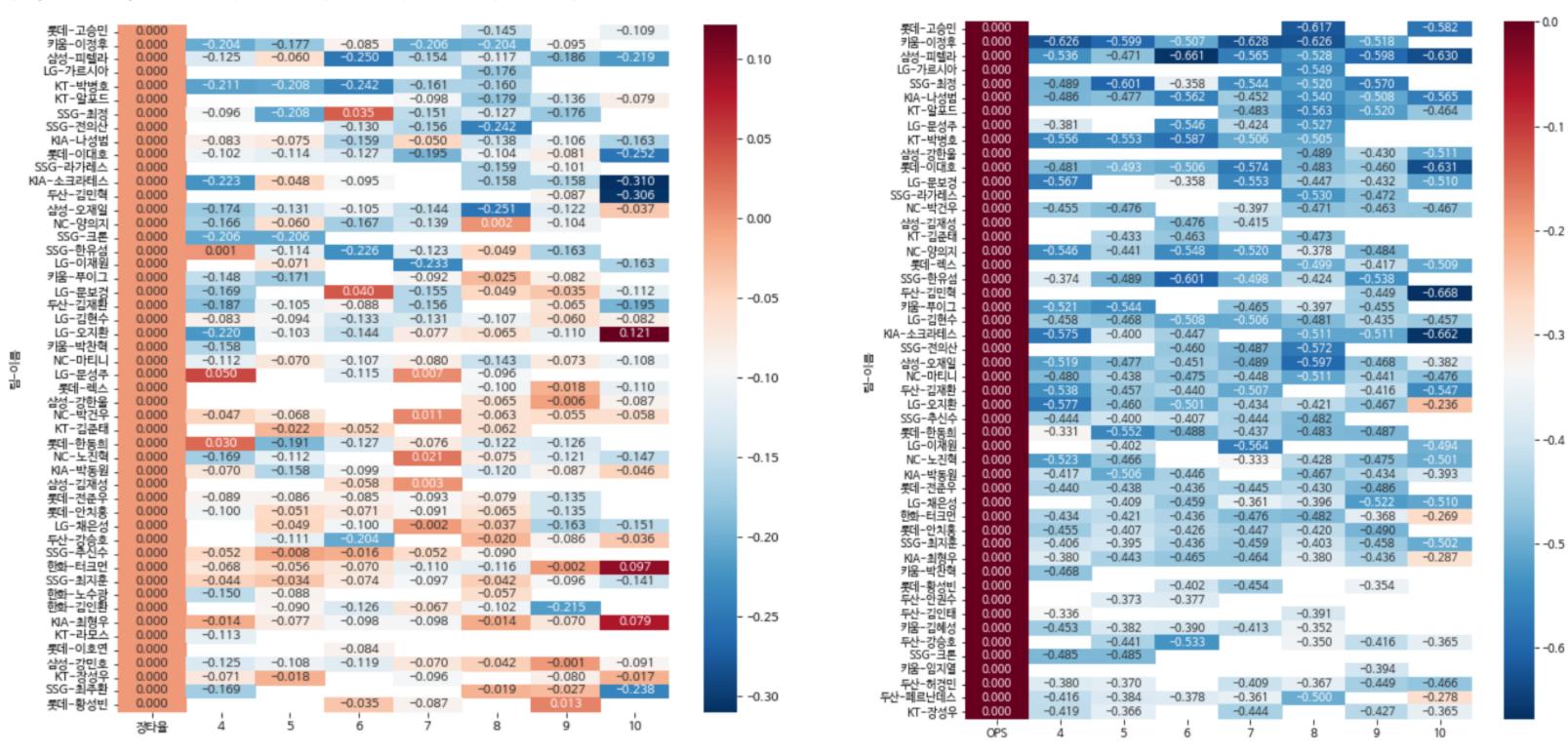


2. 야구 선수는 어떤 계절에 강해질까?

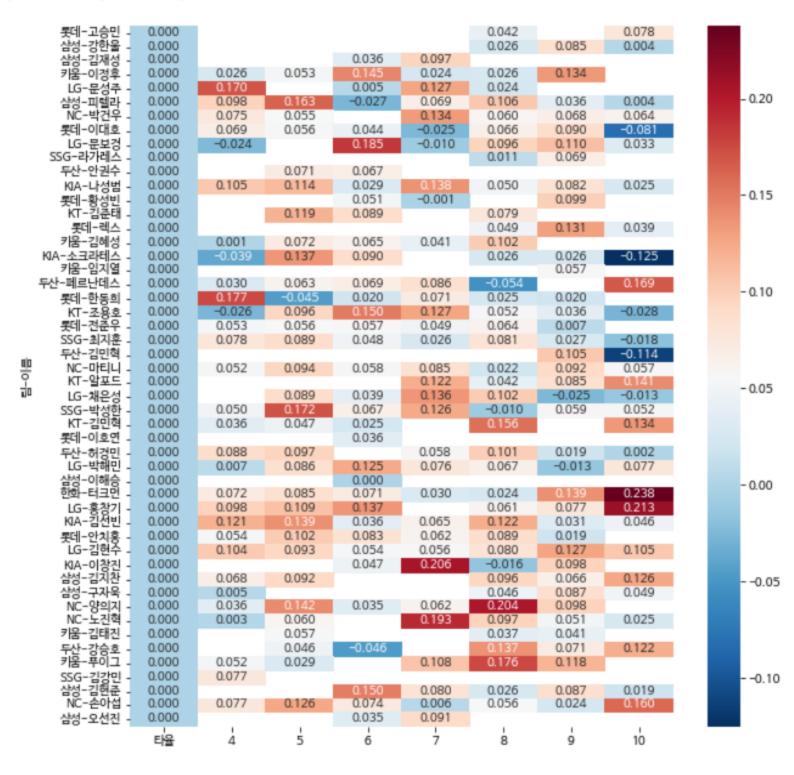
```
#월별 출루율
month_pivot = player_month_stat.pivot_table(index = ['팀', '이름'],
                                          columns = '월',
                                          values = '출루율')
month pivot = month pivot.reset index()
df = pd.merge(player_stat, month_pivot, how = 'left', on = ['팀', '이름'])
#출루율 상위 50명의 월별 데이터
df_obp = df.sort_values(by = '출루율', ascending = False).head(50)
df_obp_selected = df_obp[['팀', '이름', '출루율', 4, 5, 6, 7, 8, 9, 10]]
df_obp_selected = df_obp_selected.set_index(['팀', '이름'])
#시즌 전체 출루율과의 월별 출루율 차이
#시즌 출루율을 0으로 초기화해준다
for col in df_obp_selected.columns[1:] :
 df_obp_selected[col] = df_obp_selected[col] - df_obp_selected['출루율']
df_obp_selected['출루율'] = 0.0
fig, ax = plt.subplots(figsize=(10,10))
sns.heatmap(data = df obp_selected.head(50),
           annot = True, fmt = '.3f',
           cmap = 'RdBu r'
```

		출루율	4	5	6	ż	8	ģ	10		
	NC-노진혁 -	0.000	-0.068	-0.012		0.121	0.026	-0.021	-0.046		
	SSG-김강민 -	0.000	0.000	0.010			0.000	0.001	0.015		
	롯데-안치흥 -	0.000	-0.015	0.033	0.014	-0.007	0.020	-0.050			- 0.15
1	두산-페르난데스 -	0.000	-0.016	0.017	0.023	0.040	-0.100		0.123		0.15
	LG-오지환 -	0.000	-0.107	0.009	-0.031	0.036	0.048	0.003	0.234		
	KT-김민혁 -	0.000	-0.027	-0.016	-0.038		0.093	0.000	0.071		
	LG-채운성 -	0.000		0.028	-0.022	0.075	0.041	-0.086	-0.074		
	KIA-이창진 -	0.000			-0.027	0.132	-0.089	0.025			
	두산-허경민 -	0.000	0.017	0.027		-0.013	0.030	-0.052	-0.069		3.10
	한화-이성곤 -	0.000	0.000								0.10
	롯데-한동희 -	0.000	0.124	-0.098	-0.033	0.018	-0.028	-0.032			
	SSG-최지훈 -	0.000	0.020	0.031	-0.010	-0.032	0.023	-0.031	-0.076		
	두산-김민혁 -	0.000						0.044	-0.175		
	삼성-김현준 -	0.000			0.063	-0.006	-0.060	0.001	-0.067		
	한화-터크먼 -	0.000	-0.005	0.008	-0.007	-0.047	-0.053	0.062	0.161		
	KIA-최형우 -	0.000	0.040	-0.023	-0.044	-0.043	0.041	-0.015	0.134		0.05
	롯데-황성빈 -	0.000	0.040	0.000	0.001	-0.051	0.041	0.049	0.124		
	키움-임지열 -	0.000			0.001	0.051		0.000			
	키움-김혜성 -	0.000	-0.055	0.016	0.008	-0.015	0.046	0.000			
	NC-DELL -	0.000	-0.016	0.026	-0.010	0.017	-0.046	0.024	-0.011		
	삼성-김지찬 - NC-마티니 -	0.000	-0.016	0.008	0.010	0.017	0.012	-0.017	0.043		
	KIA-류지혁 -	0.000	0.008	0.069	-0.076	-0.018	-0.042	0.005	0.093		- 0.00
	SSG-라가레스 -	0.000	0.000	0.050	0.076	0.010	-0.034	0.025	0.003		
El u	키움-푸이그 -	0.000	-0.043	-0.066		0.013	0.081	0.023			
-B0					-0.050			-0.055	-0.040		
	LG-가르시아 - KIA-김선빈 -	0.000	0.035	0.053	-0.050	-0.021	0.000	-0.055	-0.040		
	1G-기로시아	0.000	-0.093	0.029	0.004	0.001		-0.030	-0.094		
	SSG-박성한 - KT-조용호 -	0.000	-0.027	0.029	0.084	0.049	-0.014	-0.030	-0.094		- 0.05
		0.000	-0.027	0.095	-0.010	0.049	-0.087	-0.018	-0.025		- 0.05
	LG-김현수 -	0.000	0.016	0.005	-0.035	-0.032	-0.008	0.039	0.017		
	SSG-한유섬 -	0.000	0.110	-0.005	-0.117	-0.014	0.060	-0.053	0.079		
	한화-정은원 -	0.000	-0.093	-0.015	0.106	-0.034	0.012	0.042	-0.079		
	롯데-이대호 -	0.000	0.021	0.009	-0.004	-0.072	0.019	0.042	-0.129		
	NC-양의지 -	0.000	-0.060	0.045	-0.062	-0.034	0.107	0.002	0.050		0.10
	KT-알포드 -	0.000	-0.012	0.032	0.024	0.012	-0.030	-0.001	0.056		- 0.10
	LG-홍창기 - SSG-추신수 -	0.000	-0.007	0.004	0.032	-0.012	-0.043	-0.028	U. 100		
		0.000	-0.007	0.002	0.002		-0.043	-0.028	0.108		
	SSG-최정 - 두산-안권수 -	0.000	0.030	-0.083 0.002	-0.002	-0.020	-0.002	-0.051			
	LG-문보경 -	0.000	-0.091 0.030	-0 002	0.118 0.161	-0.077 -0.026	-0.002	-0.051	-0.034		
	롯데-렉스 -	0.000	-0.001		0.110	-0.077	-0.036 0.029	0.045	-0.046		-0.13
	KIA-나성범 -	0.000	0.022	0.031	-0.054	0.056	-0.032	-0.000	-0.057		- 0.15
	두산-김인태 -	0.000	0.023	0.021	_0.0E4	0.056	-0.031	-0.000	_0.057		
	NC-박건우 -	0.000	0.003	-0.017		0.061	-0.013	-0.005	-0.008		
	KT-김준태 -	0.000	0.000	0.022	-0.008	0.051	-0.018	0.005	0.000		
	삼성-피렐라 -	0.000	0.029	0.094	-0.096	0.000	0.037	-0.032	-0.065		
	삼성-김재성 -	0.000	0.000	0.004	-0.028	0.033	0.007	0.000	0.055		- 0.20
	키움-이정후 -	0.000	-0.045	-0.018	0.074	-0.047	-0.045	0.063			- 0.20
	삼성-강한율 -	0.000					-0.027	0.032	-0.049		
	LG-문성주 -	0.000	0.083		-0.082	0.040	-0.063				
	롯데-고승민 -	0.000					-0.008		0.028		
	est a del -	0.000					0.000		0.000	_	

2. 야구 선수는 어떤 계절에 강해질까?



2. 야구 선수는 어떤 계절에 강해질까?



3. 득점과 여러 요소들의 상관관계

표준 선형 회귀, 릿지 선형 회귀, 라쏘 선형 회귀

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
df = pd.read_csv('baseball.csv', encoding = 'cp949')
from sklearn.model_selection import train_test_split
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import OneHotEncoder
num = ['홈런', '루타', '도루', '볼넷', '사구', '고4', '희타']
X = df[num]
Y = df['득점']
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size = 0.3, random_state = 0)
ct = ColumnTransformer([("scaling", StandardScaler(), num)])
ct.fit(X_train)
                                                   from sklearn.metrics import mean squared error
X_train = ct.transform(X_train)
                                                   from math import sqrt
X test = ct.transform(X test)
#표준 선형 회귀 모형
                                                   rmse = sqrt(mean_squared_error(Y_test, Y_pred))
from sklearn.linear_model import LinearRegression
                                                   print("RMSE: {:.3f}".format(rmse))
                                                   print("절편: ", np.round(lr.intercept_, 3))
lr = LinearRegression().fit(X train, Y train)
                                                   print("가중치: ", np.round(lr.coef_, 3))
Y pred = lr.predict(X test)
```

표준 선형 회귀

RMSE: 2.556

절편: 9.194

가중치: [-0.146 3.241 1.004 0.738

0.281 -0.093 0.214]

3. 득점과 여러 요소들의 상관관계

표준 선형 회귀, 릿지 선형 회귀, 라쏘 선형 회귀

릿지 선형 회귀

RMSE: 2.556

절편: 9.194

가중치: [-0.133 3.221 1.006 0.74 0.28 -0.09 0.213]

라쏘 선형 회귀

RMSE: 2.556

절편: 9.194

가중치: [-0.143 3.238 1.004 0.737 0.28 -0.091 0.213]

연구결과 활용방안









- 1. 스포츠는 변수가 많다.
- 2. 타율, 출루율, 장타율, OPS 외에도 다양한 지표들이 존재한다.
- 3. 수비 관련 지표로도 연구가 가능하다.

연구 한계 & 후속 연구 제안