EPL EDA 프로젝트

잉글랜드 프리미어 리그의 2014년부터 2022년까지
선수 연봉 및 스탯 자료 분석 프로젝트

팀명: EDA 챔피언스⚽
팀원: 이인서, 이도형, 김동억, 권도형
2023. 9. 8

content

1. 데이터 소개

데이터 소개

잉글랜드 프리미어 리그 데이터

❑ 2014년부터 2022년까지 연봉과 선수 스탯에 대한 자료

❑ 연봉 자료 정보

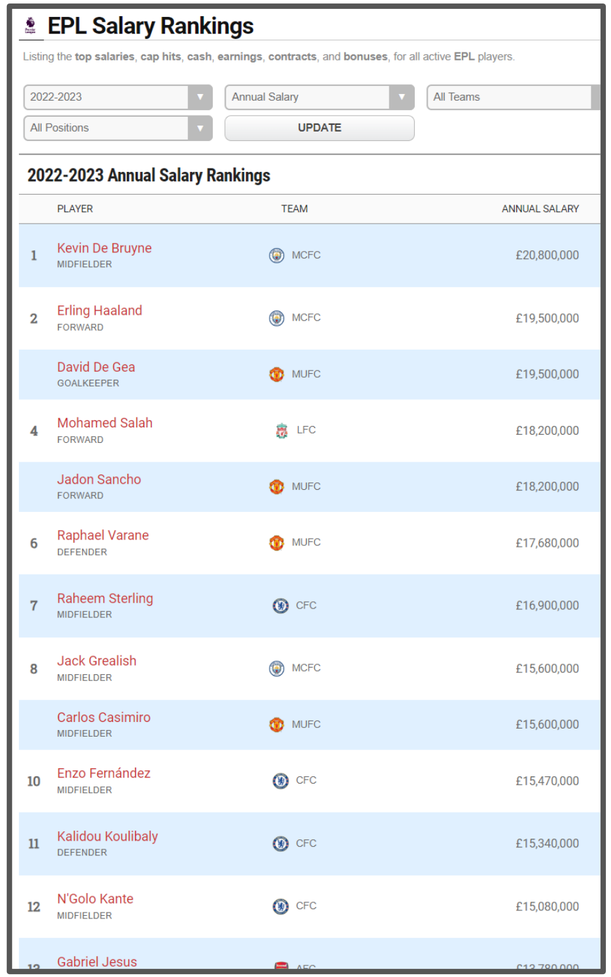
- ❑ Weekly Salary : 선수의 주급(영국 축구 선수는 주급 체계)
- ❑ Base Salary : 선수의 기본 총 급여(흔히 말하는 연봉)
- ❑ ADJ Salary : 조정된 선수의 기본 총 급여

❑ 선수 스탯 정보

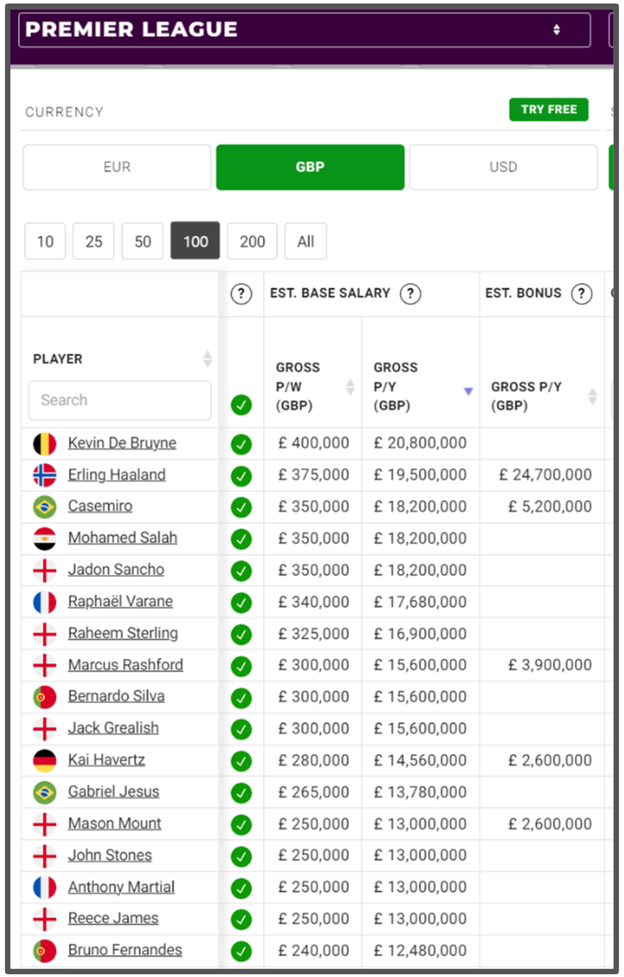
- ❑ Apps : 총 출전 횟수
- ❑ Mins : 총 출전 시간(분)
- ❑ Rating : 시즌 평점
- ❑ SpG : 경기당 슈팅 횟수
- ❑ KeyP : 경기당 키패스 횟수
- ❑ Fouled : 경기당 파울 당한 횟수
- ❑ Off : 경기당 오프사이드에 걸린 횟수
- ❑ Disp : 경기당 드리블 실수로 (공 소유권 잃음)
- ❑ Drb_Off : 경기당 드리블 성공 횟수(공격지표)
- ❑ Drb_Def : 경기당 드리블을 당한 횟수(수비지표)
- ❑ UnsTch : 공 컨트롤 실수로 (공 소유권 잃음)
- ❑ Tackles : 경기당 태클 수
- ❑ Inter : 경기당 가로채기 수
- ❑ Fouls : 경기당 파울 한 횟수
- ❑ Offsides : 경기당 오프사이드 트랩 성공 횟수
- ❑ Clear : 경기당 공을 걷어 낸 횟수
- ❑ Blocks : 경기당 가로막은 횟수
- ❑ **G : 한 시즌의 총 득점 수**
- ❑ **A : 한 시즌의 총 도움 수**
- ❑ **xG : 기대 득점 값(전 시즌 기준으로 계산 됨)**
- ❑ **xA : 기대 도움 값(전 시즌 기준으로 계산 됨)**
- ❑ AvgP : 경기당 패스 횟수
- ❑ PS% : 패스 성공률
- ❑ NPG : 페널티킥 없는 득점 수
- ❑ NPxG : 패널티킥 없는 기대 득점 값
- ❑ **xG Chain : 슈팅까지 연결된 체인에 관여한 모든 선수에게 주는 기댓값**
- ❑ **xG Buildup : 체인에서 슈팅과 키패스에 관여하지 않은 선수에게 주는 값**
- ❑ xG 90 : 90분당 xG값
- ❑ NPxG 90 : 90분당 NPxG값
- ❑ xA 90 : 90분당 xA
- ❑ xG90 + xA90 : 90분당 공격포인트에 대한 기댓값
- ❑ NPxG90 + xA90 : 90분당 페널티킥 없는 공격포인트 기댓값(더 객관적 지표)
- ❑ xGChain90 : 90분당 xG Chain값
- ❑ xGBuildup90 : 90분당 xG Buildup값

데이터 수집 시작

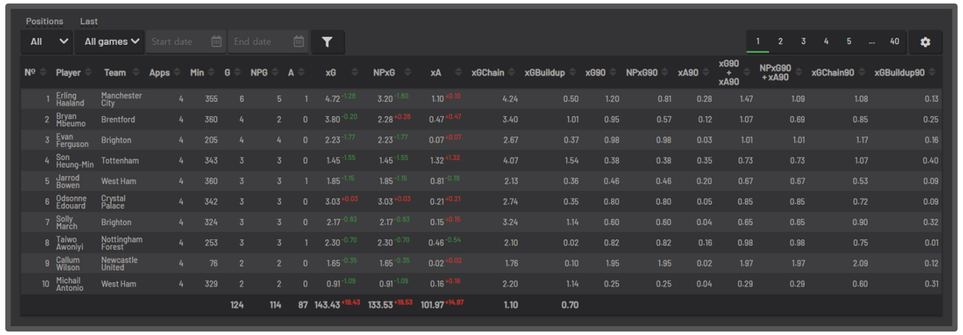
총 4개의 사이트 분담해서 크롤링 하기!

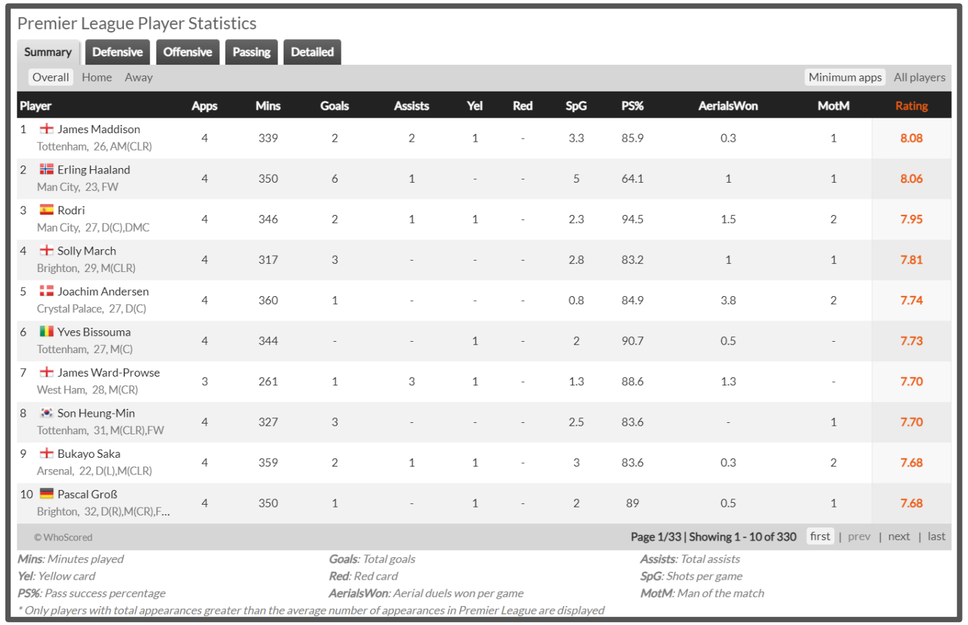


spotrac



capology





understat(상), 1xbet(하)

understat 크롤링

[담당:김동억]

Positions Last

All | All games | Start date | End date | 1 2 3 4 5 ... 40

№	Player	Team	Apps	Min	G	NPG	A	xG	NPxG	xA	xGChain	xGBuildup	xG90	NPxG90	xA90	xG90 + xA90	NPxG90 +xA90	xGChain90	xGBuildup90
1	Erling Haaland	Manchester City	4	355	6	5	1	4.72 -1.28	3.20 -1.80	1.10 +0.10	4.24	0.50	1.20	0.81	0.28	1.47	1.09	1.08	0.13
2	Bryan Mbeumo	Brentford	4	360	4	2	0	3.80 -0.20	2.28 +0.28	0.47 +0.47	3.40	1.01	0.95	0.57	0.12	1.07	0.69	0.85	0.25
3	Evan Ferguson	Brighton	4	205	4	4	0	2.23 -1.77	2.23 -1.77	0.07 +0.07	2.67	0.37	0.98	0.98	0.03	1.01	1.01	1.17	0.16
4	Son Heung-Min	Tottenham	4	343	3	3	0	1.45 -1.55	1.45 -1.55	1.32 +1.32	4.07	1.54	0.38	0.38	0.35	0.73	0.73	1.07	0.40
5	Jarrod Bowen	West Ham	4	360	3	3	1	1.85 -1.15	1.85 -1.15	0.81 -0.19	2.13	0.36	0.46	0.46	0.20	0.67	0.67	0.53	0.09
6	Odsonne Edouard	Crystal Palace	4	342	3	3	0	3.03 +0.03	3.03 +0.03	0.21 +0.21	2.74	0.35	0.80	0.80	0.05	0.85	0.85	0.72	0.09
7	Solly March	Brighton	4	324	3	3	0	2.17 -0.83	2.17 -0.83	0.15 +0.15	3.24	1.14	0.60	0.60	0.04	0.65	0.65	0.90	0.32
8	Taiwo Awoniyi	Nottingham Forest	4	253	3	3	1	2.30 -0.70	2.30 -0.70	0.46 -0.54	2.10	0.02	0.82	0.82	0.16	0.98	0.98	0.75	0.01
9	Callum Wilson	Newcastle United	4	76	2	2	0	1.65 -0.35	1.65 -0.35	0.02 +0.02	1.76	0.10	1.95	1.95	0.02	1.97	1.97	2.09	0.12
10	Michail Antonio	West Ham	4	329	2	2	0	0.91 -1.09	0.91 -1.09	0.16 +0.16	2.20	1.14	0.25	0.25	0.04	0.29	0.29	0.60	0.31
					124	114	87	143.43 +18.43	133.53 +19.53	101.97 +14.97	1.10	0.70							

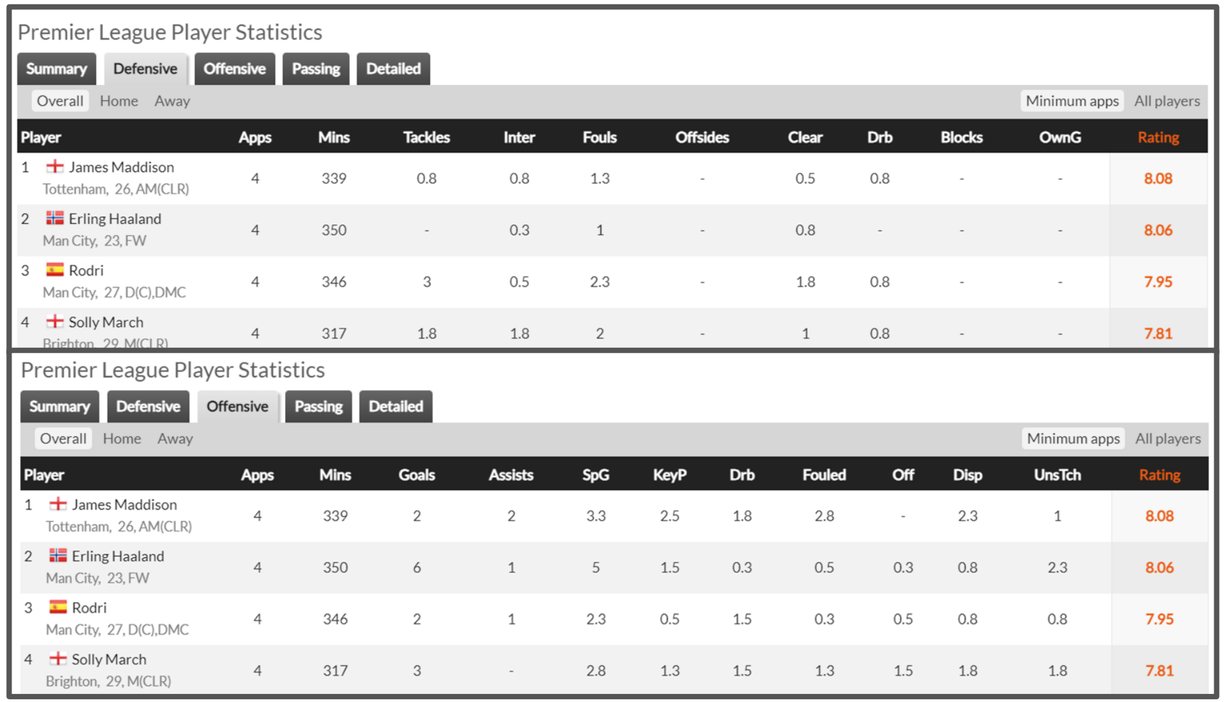
```
No,Name,Team,Apps,Min,G,NPG,A,xG,NPxG,xA,xGChain,xGBuildup,xG90,NPxG90,xA90,xG90+xA90,NPxG90+xA90,xGChain90,xGBuildup9
619,Sergio Aguero,Manchester City,33,2551,26,21,8,25.27,20.7,5.57,27.81,6.88,0.89,0.73,0.2,1.09,0.93,0.98,0.24,2014
647,Harry Kane,Tottenham,34,2589,21,19,4,17.16,14.87,3.92,16.49,5.55,0.6,0.52,0.14,0.73,0.65,0.57,0.19,2014
802,Diego Costa,Chelsea,26,2111,20,19,3,15.22,14.46,4.55,21.37,5.28,0.65,0.62,0.19,0.84,0.81,0.91,0.22,2014
848,Charlie Austin,Queens Park Rangers,35,3078,18,15,5,17.88,14.08,2.55,13.72,3.04,0.52,0.41,0.07,0.6,0.49,0.4,0.09,20
498,Alexis Sanchez,Arsenal,35,2967,16,16,8,13.45,12.69,8.49,27.16,10.74,0.41,0.38,0.26,0.67,0.64,0.82,0.33,2014
502,Olivier Giroud,Arsenal,27,1871,14,14,3,8.85,8.85,3.86,13.64,4.13,0.43,0.43,0.19,0.61,0.61,0.66,0.2,2014
701,Eden Hazard,Chelsea,38,3389,14,11,9,12.02,8.97,11.24,31.84,19.48,0.32,0.24,0.3,0.62,0.54,0.85,0.52,2014
811,Saido Berahino,West Bromwich Albion,38,2940,14,10,1,13.84,10.8,1.96,11.91,2.44,0.42,0.33,0.06,0.48,0.39,0.36,0.07,
606,Christian Benteke,Aston Villa,29,2380,13,12,2,8.46,7.69,2.92,8.73,2.2,0.32,0.29,0.11,0.43,0.4,0.33,0.08,2014
617,David Silva,Manchester City,32,2682,12,12,7,9.1,9.1,10.39,29.14,16.63,0.31,0.31,0.35,0.65,0.65,0.98,0.56,2014
629,Wayne Rooney,Manchester United,33,2887,12,11,5,8.9,7.38,3.98,15.44,7.3,0.28,0.23,0.12,0.4,0.35,0.48,0.23,2014
841,Graziano Pelle,Southampton,38,3291,12,12,2,18.62,18.62,4.48,24.79,5.6,0.51,0.51,0.12,0.63,0.63,0.68,0.15,2014
622,Wilfried Bony,Manchester City,30,1604,11,10,3,10.45,9.69,3.51,14.25,4.41,0.59,0.54,0.2,0.78,0.74,0.8,0.25,2014
```

❑ 사이트 정보

- ❑ 사이트명 : understat
- ❑ 사이트 특징 : 선수들의 2014~2022 스탯 제공
- ❑ 전체 데이터 수 : 4755
- ❑ 결측치 : -
- ❑ 사용피쳐 : Name, Team, Apps, Min, G, NPG,
  A, xG, NPxG, xA, xGChain, xG Buildup, xG90, NPxG90, xA90, xG90+xA90, NPxG90+xA90, xGChain90, xGBuildup90
- ❑ 주의사항 :
  - ❑ 설정을 통해 수집하고자 하는 스탯을 설정해야만 크롤링 가능
  - ❑ 화면에 10명의 선수가 표시되고 페이징을 통해 나머지 선수들이 표시됨
- ❑ 수집결과 : uderstat_(2014~2022).csv

1xbet 크롤링

[수비지표 담당:이도형, 공격지표 담당:이인서]



Premier League Player Statistics

Summary | Defensive | Offensive | Passing | Detailed

Overall | Home | Away — Minimum apps | All players

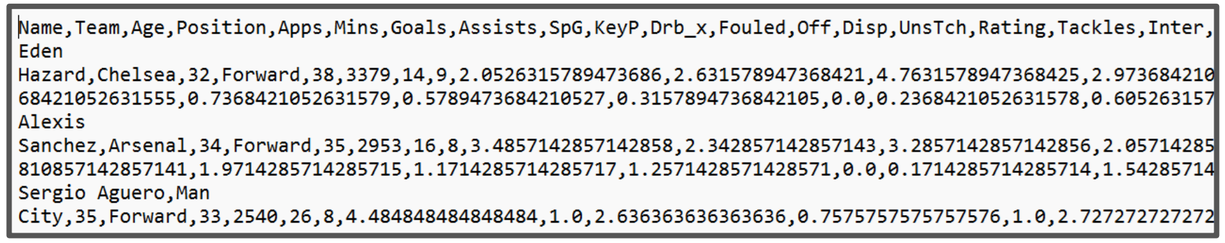
	Player	Apps	Mins	Tackles	Inter	Fouls	Offsides	Clear	Drb	Blocks	OwnG	Rating
1	James Maddison Tottenham, 26, AM(CLR)	4	339	0.8	0.8	1.3	-	0.5	0.8	-	-	8.08
2	Erling Haaland Man City, 23, FW	4	350	-	0.3	1	-	0.8	-	-	-	8.06
3	Rodri Man City, 27, D(C),DMC	4	346	3	0.5	2.3	-	1.8	0.8	-	-	7.95
4	Solly March Brighton, 29, M(CLR)	4	317	1.8	1.8	2	-	1	0.8	-	-	7.81

Premier League Player Statistics

Summary | Defensive | Offensive | Passing | Detailed

Overall | Home | Away — Minimum apps | All players

	Player	Apps	Mins	Goals	Assists	SpG	KeyP	Drb	Fouled	Off	Disp	UnsTch	Rating
1	James Maddison Tottenham, 26, AM(CLR)	4	339	2	2	3.3	2.5	1.8	2.8	-	2.3	1	8.08
2	Erling Haaland Man City, 23, FW	4	350	6	1	5	1.5	0.3	0.5	0.3	0.8	2.3	8.06
3	Rodri Man City, 27, D(C),DMC	4	346	2	1	2.3	0.5	1.5	0.3	0.5	0.8	0.8	7.95
4	Solly March Brighton, 29, M(CLR)	4	317	3	-	2.8	1.3	1.5	1.3	1.5	1.8	1.8	7.81



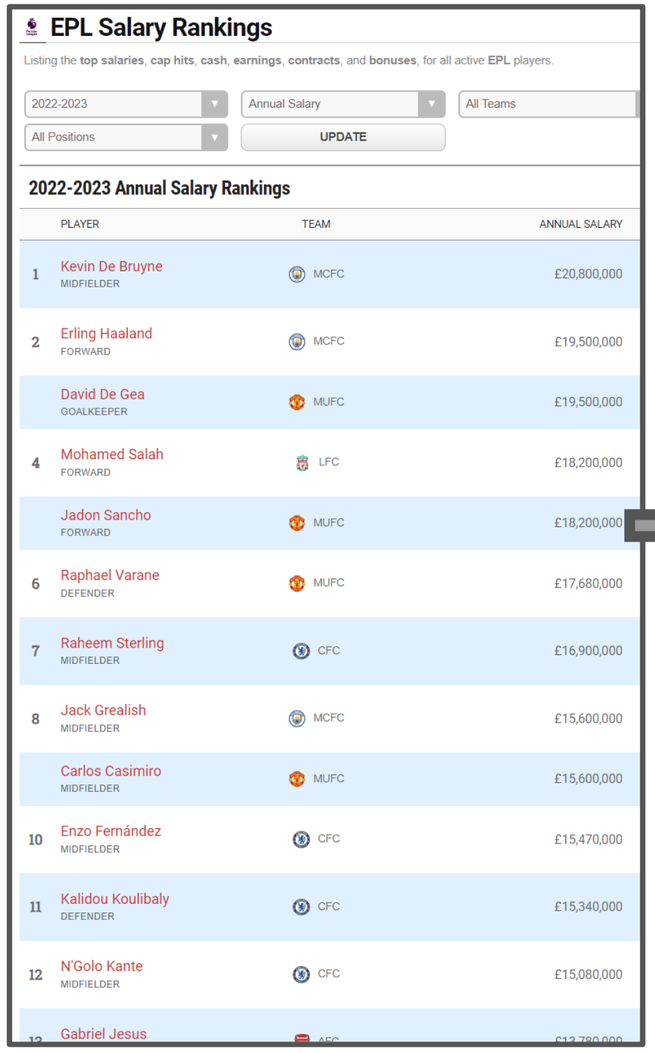
```
Name,Team,Age,Position,Apps,Mins,Goals,Assists,SpG,KeyP,Drb_x,Fouled,Off,Disp,UnsTch,Rating,Tackles,Inter,
Eden
Hazard,Chelsea,32,Forward,38,3379,14,9,2.0526315789473686,2.631578947368421,4.7631578947368425,2.973684210
68421052631555,0.7368421052631579,0.5789473684210527,0.3157894736842105,0.0,0.2368421052631578,0.605263157
Alexis
Sanchez,Arsenal,34,Forward,35,2953,16,8,3.4857142857142858,2.342857142857143,3.2857142857142856,2.05714285
810857142857141,1.9714285714285715,1.1714285714285717,1.2571428571428571,0.0,0.1714285714285714,1.54285714
Sergio Aguero,Man
City,35,Forward,33,2540,26,8,4.484848484848484,1.0,2.636363636363636,0.7575757575757576,1.0,2.727272727272
```

❑ 사이트 정보

- ❑ 사이트명 : 1xbet
- ❑ 사이트 특징 : 선수들의 2014~2022 수비, 공격, 패싱 지표를 카테고리로 제공
- ❑ 전체 데이터 수 : 4756
- ❑ 결측치 : -
- ❑ 사용피쳐 :Name, Apps, Goals, Assists, SpG, KeyP, Drb_Off, Fouled, Off, Disp, UnsTch, Rating, Tackles, Inter,Fouls, Offsides,Clear, Drb_Def, Blocks, AvgP, PS%
- ❑ 주의사항 :
  - ❑ 셀레늄을 사용해서 크롤링하려고 했는데 자꾸 광고가 뜨는 오류
  - ❑ html 정보가 제대로 업데이트 되지 않는 오류
  - ❑ Request URL을 통해 자료 수집
- ❑ 수집결과 :
  1xbet_def(2014~2022).csv
  1xbet_off(2014~2022).csv

spotrac 크롤링

[담당:권도형]



EPL Salary Rankings

Listing the **top salaries**, **cap hits**, **cash**, **earnings**, **contracts**, and **bonuses**, for all active **EPL** players.

2022-2023 | Annual Salary | All Teams | All Positions | UPDATE

2022-2023 Annual Salary Rankings

	PLAYER	TEAM	ANNUAL SALARY
1	Kevin De Bruyne MIDFIELDER	MCFC	£20,800,000
2	Erling Haaland FORWARD	MCFC	£19,500,000
	David De Gea GOALKEEPER	MUFC	£19,500,000
4	Mohamed Salah FORWARD	LFC	£18,200,000
	Jadon Sancho FORWARD	MUFC	£18,200,000
6	Raphael Varane DEFENDER	MUFC	£17,680,000
7	Raheem Sterling MIDFIELDER	CFC	£16,900,000
8	Jack Grealish MIDFIELDER	MCFC	£15,600,000
	Carlos Casimiro MIDFIELDER	MUFC	£15,600,000
10	Enzo Fernández MIDFIELDER	CFC	£15,470,000
11	Kalidou Koulibaly DEFENDER	CFC	£15,340,000
12	N'Golo Kante MIDFIELDER	CFC	£15,080,000
	Gabriel Jesus	AFC	£13,780,000

```
Rank,Name,Position,Team,Weekly Salary,year
1.0,Fernando Torres,Forward,CFC,340362,2014
2.0,Wayne Rooney,Forward,MUFC,300000,2014
3.0,Sergio Aguero,Forward,MCFC,220000,2014
3.0,Yaya Toure,Midfielder,MCFC,220000,2014
5.0,Diego Costa,Forward,CFC,185000,2014
6.0,David Silva,Midfielder,MCFC,160000,2014
6.0,Juan Mata,Midfielder,MUFC,160000,2014
8.0,Lukas Podolski,Forward,AFC,157769,2014
9.0,Cesc Fabregas,Midfielder,CFC,156000,2014
10.0,Andre Schurrle,Forward,CFC,152623,2014
11.0,Oscar,Midfielder,CFC,152255,2014
12.0,Eden Hazard,Midfielder,CFC,150000,2014
12.0,Vincent Kompany,Defender,MCFC,150000,201
12.0,Samir Nasri,Forward,MCFC,150000,2014
15.0,Mesut Ozil,Midfielder,AFC,140000,2014
15.0,Alexis Sanchez,Forward,AFC,140000,2014
17.0,Darren Bent,Forward,AVFC,137769,2014
18.0,Demba Ba,Forward,CFC,137276,2014
19.0,Daniel Sturridge,Forward,LFC,130000,2014
20.0,Mikel Arteta,Midfielder,AFC,125769,2014
21.0,Danny Welbeck,Forward,AFC,125000,2014
22.0,Petr Cech,Goalkeeper,CFC,120000,2014
22.0,Thibaut Courtois,Goalkeeper,CFC,120000,2
24.0,Joe Hart,Goalkeeper,MCFC,116000,2014
25.0,Ander Herrera,Midfielder,MUFC,115000,201
26.0,Aaron Ramsey,Midfielder,AFC,110000,2014
26.0,Marcos Rojo,Defender,MUFC,110000,2014
26.0,Luke Shaw,Defender,MUFC,110000,2014
29.0,Wojciech Szczesny,Goalkeeper,AFC,100000,
29.0,Theo Walcott,Forward,AFC,100000,2014
29.0,Wilfried Bony,Forward,MCFC,100000,2014
29.0,Fernandinho Luis Roza,Midfielder,MCFC,10
29.0,Eliaquim Mangala,Defender,MCFC,100000,20
29.0,Fernando Reges,Midfielder,MCFC,100000,20
```

❑ 사이트 정보

- ❑ 사이트명 : spotrac
- ❑ 사이트 특징 : 선수들의 2014~2022 연봉 제공
- ❑ 전체 데이터 수 : 4803
- ❑ 결측치 : Rank - 1252/4803
- ❑ 사용피쳐 : Name, Position, Team, Weekly Salary(capology로 대체됨)
- ❑ 주의사항 :
  - ❑ 같은 Rank의 경우 첫 선수만 Rank가 표시되고 이 후 나오는 선수는 결측치로 설정됨
- ❑ 수집결과 : salary_(2014~2022).csv

content

전처리 시작

크롤링한 데이터 모두 하나로 통합하는 과정

- ❏ 회의 내용
- ❏ 선수 이름 정리
  - ❏ **영어 알파벳으로 통합**(스페인어, 덴마크어 등)
  - ❏ 필요없는 값 제거(공백, 임대 제거)
- ❏ **이적선수 스탯 통합**
  - ❏ 이적으로 인해 동일한 해에 스탯이 분리된 경우 경기수를 반영하여 스탯 통합
- ❏ **선수 프로필 테이블 생성**
  - ❏ 데이터 통합 시 기준이 되는 테이블
  - ❏ 연도, 선수이름, 팀, 포지션의 데이터로 구성
- ❏ 연봉 정보 수정
  - ❏ 화폐 기호 및 콤마 제거
  - ❏ int로 변환

이름 알파벳 통합

```
페데리코 페르난데스,Swansea,34,Defender,28,2475,1.928571428
당신의 디아프라,West Ham,33,Forward,23,1761,0.8695652173913
마이클 도슨,Hull,39,Defender,28,2434,1.4285714285714286,1.6
조이 바튼,QPR,40,Midfielder,28,2351,2.9642857142857144,1.64
기상윙,Swansea,34,Midfielder,33,2690,1.3636363636363635,1.
오레스 오코레,Aston Villa,31,Defender,23,2009,2.217391304347
알몸의 염소,Man City,40,Defender,9,740,2.2222222222222223,0
제임스 매카시,Everton,32,Midfielder,28,2404,2.428571428571
라이언 쇼크로스,Stoke,35,Defender,32,2832,1.28125,1.5625,0.
크리스티안 에릭센,Tottenham,31,Midfielder,38,3142,1.342105
루카스 파비안스키,Swansea,38,Goalkeeper,37,3308,0.027027027
벤 미,Burnley,33,Defender,33,2888,1.9090909090909087,2.030
패트릭 반 안홀트,Sunderland,33,Defender,28,2306,2.71428571
크리스티안 벤테케,Aston Villa,32,Forward,29,2380,0.37931034
파비안 델프,Aston Villa,33,Midfielder,28,2434,2.10714285714
리스 버크,West Ham,26,Defender,5,415,1.2,1.8,0.6,0.8,10.4,
아론 램지,Arsenal,32,Midfielder,29,2010,2.0,1.1034482758620
리차드 던,QPR,43,Defender,23,1874,2.6956521739130435,1.913
새미 아메오비,Newcastle,31,Forward,25,1423,2.0,1.36,1.36,0.
토비 알더베이럴트,Southampton,34,Defender,26,2263,1.6923076
사디오 마네,Southampton,31,Forward,30,2135,1.3,0.5,1.133333
메르테자커 당,Arsenal,38,Defender,35,3150,0.7714285714285714
에릭 라믈라,Tottenham,31,Midfielder,33,2302,2.3636363636363
엘리아퀌 망갈라,Man City,32,Defender,25,2189,1.72,1.6,0.84,
데이비드 오스피나,Arsenal,34,Goalkeeper,18,1620,0.05555555
얀 베르통언,Tottenham,36,Defender,32,2810,1.59375,2.1875,0.
마이클 캐릭,Man Utd,42,Midfielder,18,1457,1.388888888888888
```

1차 변환 시도 실패 : 구글 트랜스 이용해서 한글로 번역 시도했는데 이상한 단어로 번역 (기상윙, 알몸의 염소 등)

```python
import pandas as pd

special_characters = {
          'À': 'A', 'Á': 'A','Â': 'A','Ã': 'A','Ä': 'A','Å': 'A','Ā': 'A','Ă': 'A','Ą': 'A','Æ': 'Ae',
          'Ç': 'C','È': 'E','É': 'E','Ê': 'E','Ë': 'E','Ì': 'I','Í': 'I','Î': 'I','Ï': 'I',
          'Ñ': 'N','Ò': 'O','Ó': 'O','Ô': 'O','Õ': 'O','Ö': 'O','Ø': 'O','Ù': 'U',
          'Ú': 'U','Û': 'U','Ü': 'U','ß': 'ss','à': 'a','á': 'a','â': 'a',
          'ã': 'a','ä': 'a','å': 'a','æ': 'ae','ç': 'c','è': 'e','é': 'e',
          'ê': 'e','ë': 'e','ì': 'i','í': 'i','î': 'i','ï': 'i','ñ': 'n','ò': 'o','ó': 'o',
          'ô': 'o','õ': 'o','ö': 'o','ø': 'o','ù': 'u','ú': 'u','û': 'u','ü': 'u',
          'ý': 'y','Ć': 'C','ć': 'c','Č': 'C','č': 'c','Đ': 'Dj','đ': 'dj','Ğ': 'G',
          'ğ': 'g','İ': 'I','ı': 'i','Ş': 'S','ş': 's','Š': 'S','š': 's','Ÿ': 'Y','Ž': 'Z',
          'ž': 'z'}

def trans_and_convert(name):
    english_alphabet_name = ''.join([special_characters[char] if char in special_characters else char for char in name])
    return english_alphabet_name


years_range = list(range(2014, 2023))


def alphbet_convert(season):
    file_path = fr"C:\Users\LEGION\Downloads\EDA 프로젝트\1차 자료\1xbet_offensive\1xbet_offensive_{season}.csv"
    file_name = pd.read_csv(file_path)

    # Translate the "Name" column
    for index, row in file_name.iterrows():
        name_en = trans_and_convert(row["Name"])
        file_name.at[index, "Name"] = name_en
    # Save the translated DataFrame
```

2차 변환 시도 성공 : 알파벳이 다른 나라를 검색해서 다른 부분만 하드코딩으로 변경

중복 선수 데이터 통합

49	Theo Walcott	Everton	34	Forward	14	1153	3	3	1.357143
494	Theo Walcott	Arsenal	34	Forward	6	64	0	0	0.000000

1xbet 사이트의 선수들은 확인하다보니 한 시즌에 이적을 했던 선수들은 프로필이 2개로 확인됨

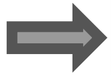
```
for d in range(0, len(duplicated_df), 2):
    data1, data2 = duplicated_df.iloc[d], duplicated_df.iloc[d+1]
    index1,index2 = data1.name, data2.name
    apps1, apps2 = data1['Apps'], data2['Apps']
    total_apps = apps1 + apps2
    stats = []
    for s1, s2 in zip(data1, data2):
        stats.append((s1*apps1+s2*apps2)/total_apps if isinstance(s1, float) else s1+s2 if isinstance(s1, np.int64) else s1)
    stats[2] = data1['Age']
    stats[-1] = data1['year']
    stats_df = stats_df.drop(index1).drop(index2)
    stats_df = pd.concat([stats_df, pd.DataFrame([stats], columns=stats_df.columns)])
stats_df.info()
```

경기 수와 출전시간은 그냥 더해주고 나머지 피쳐들은 더해준 후 경기 수로 나눠줘서 통합 프로필 완성!

선수 테이블 생성

	index	Year	Name	Age	Team	Position	Player Id
62	62	2014	Danny Welbeck	32	Arsenal	Forward	39308
420	420	2014	Danny Welbeck	32	Man Utd	Forward	39308
643	643	2015	Danny Welbeck	32	Arsenal	Forward	39308
1431	1431	2016	Danny Welbeck	32	Arsenal	Forward	39308
1998	1998	2017	Danny Welbeck	32	Arsenal	Forward	39308
2558	2558	2018	Danny Welbeck	32	Arsenal	Forward	39308
3070	3070	2019	Danny Welbeck	32	Watford	Forward	39308
3443	3443	2020	Danny Welbeck	32	Brighton	Forward	39308
3960	3960	2021	Danny Welbeck	32	Brighton	Forward	39308
4419	4419	2022	Danny Welbeck	32	Brighton	Forward	39308

	index	Year	Name	Age	Team	Position	Player Id
62	62	2014	Danny Welbeck	32	Arsenal	Forward	39308
643	643	2015	Danny Welbeck	32	Arsenal	Forward	39308
1431	1431	2016	Danny Welbeck	32	Arsenal	Forward	39308
1998	1998	2017	Danny Welbeck	32	Arsenal	Forward	39308
2558	2558	2018	Danny Welbeck	32	Arsenal	Forward	39308
3070	3070	2019	Danny Welbeck	32	Watford	Forward	39308
3443	3443	2020	Danny Welbeck	32	Brighton	Forward	39308
3960	3960	2021	Danny Welbeck	32	Brighton	Forward	39308
4419	4419	2022	Danny Welbeck	32	Brighton	Forward	39308



```
,No.,year,Name,Age,Team,Position
0,450,2014,Eden Hazard,23,Chelsea,Forward
1,74,2014,Alexis Sanchez,25,Arsenal,Forward
2,1508,2014,Sergio Aguero,26,Man City,Forward
3,261,2014,Cesc Fabregas,27,Chelsea,Midfielder
4,1480,2014,Santi Cazorla,29,Arsenal,Midfielder
5,1175,2014,Mesut Ozil,25,Arsenal,Midfielder
6,1201,2014,Mile Jedinak,30,Crystal Palace,Midfielder
7,385,2014,David Silva,28,Man City,Midfielder
8,1351,2014,Phil Jones,22,Man Utd,Defender
9,1217,2014,Morgan Schneiderlin,24,Southampton,Midfielder
10,1259,2014,Nemanja Matic,26,Chelsea,Midfielder
11,544,2014,Francis Coquelin,23,Arsenal,Midfielder
12,1666,2014,Victor Moses,23,Stoke,Midfielder
13,416,2014,Diego Costa,25,Chelsea,Forward
14,1409,2014,Robert Huth,30,Leicester,Defender
15,1720,2014,Yaya Toure,31,Man City,Midfielder
16,718,2014,James Tomkins,25,West Ham,Defender
17,978,2014,Laurent Koscielny,28,Arsenal,Defender
18,519,2014,Federico Fazio,27,Tottenham,Defender
19,560,2014,Gael Clichy,29,Man City,Defender
20,559,2014,Gabriel Paulista,23,Arsenal,Defender
21,197,2014,Boaz Myhill,31,WBA,Goalkeeper
22,284,2014,Chris Smalling,24,Man Utd,Defender
23,824,2014,Jonas Gutierrez,31,Newcastle,Midfielder
24,1354,2014,Philippe Coutinho,22,Liverpool,Midfielder
25,1230,2014,Nacho Monreal,28,Arsenal,Defender
26,93,2014,Ander Herrera,25,Man Utd,Midfielder
27,329,2014,Curtis Davies,29,Hull,Defender
28,54,2014,Aleksandar Kolarov,28,Man City,Defender
29,903,2014,Jussi Jaaskelainen,39,West Ham,Goalkeeper
30,622,2014,Harry Kane,21,Tottenham,Forward
31,115,2014,Andy Carroll,25,West Ham,Forward
32,1682,2014,Wayne Rooney,28,Man Utd,Forward
33,1366,2014,Raheem Sterling,19,Liverpool,Forward
```

통합을 위한 기준 테이블 만들기

- 동일 시즌에 2개의 데이터가 있는 선수들을 통합
- 선수들의 년도 별 나이를 계산하여 입력
- 출생년도 추가
- 고유 번호 추가

결측률 확인(1)

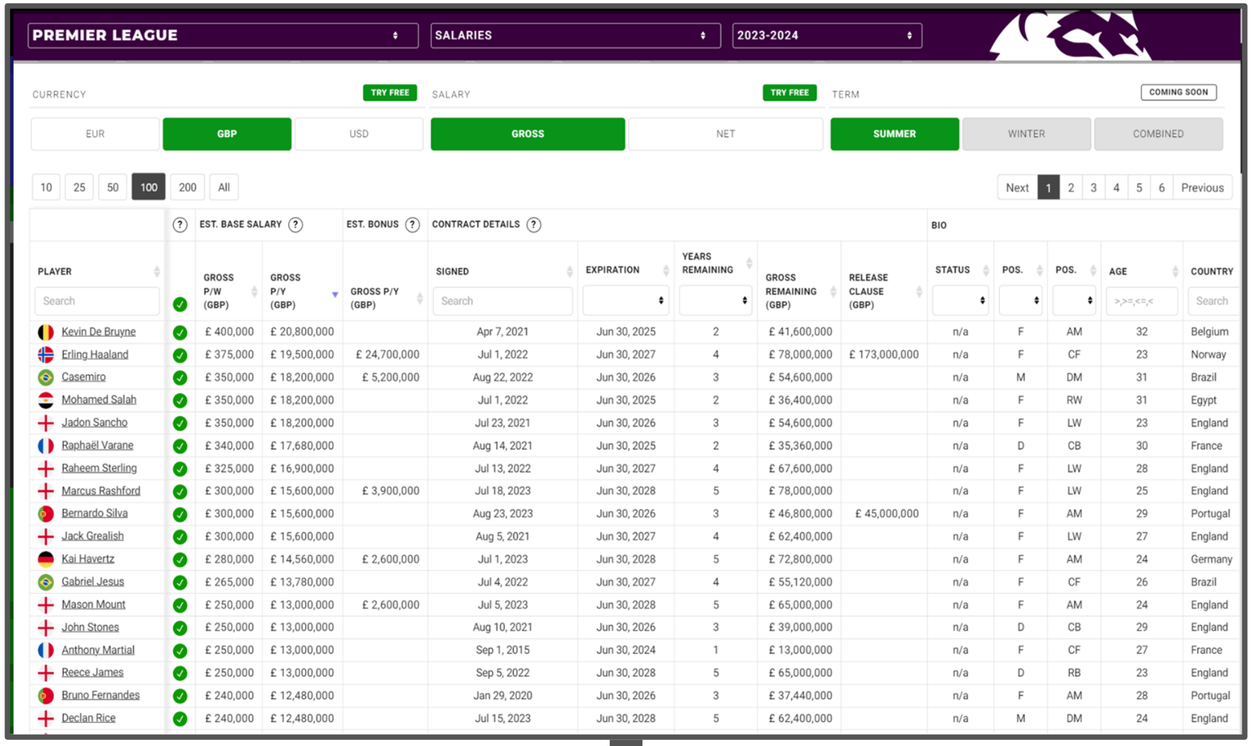
1차 완성 된 데이터 통합 결과!

```
eda_df.isnull().mean()
```

```
Unnamed: 0          0.000000
No.                 0.000000
Year                0.000000
Name                0.000000
Age                 0.000000
Team_x              0.000000
Position_x          0.000000
year                0.000000
Position_y          0.342221
Team_y              0.342221
Weekly Salary       0.342221
dtype: float64
```

- ❏ 통합 과정
  - ❏ Player 테이블과 spotrac의 연봉정보 결합
- ❏ 문제점
  - ❏ **결측률 약 34퍼센트**
- ❏ 해결 방안
  - ❏ 새로운 연봉 사이트 서치
    - ❏ capology 사이트 크롤링

추가 사이트 크롤링



PREMIER LEAGUE | SALARIES | 2023-2024

CURRENCY (TRY FREE) EUR | GBP | USD — SALARY (TRY FREE) GROSS | NET — TERM (COMING SOON) SUMMER | WINTER | COMBINED

10 | 25 | 50 | 100 | 200 | All — Next 1 2 3 4 5 6 Previous

PLAYER		EST. BASE SALARY: GROSS P/W (GBP)	EST. BASE SALARY: GROSS P/Y (GBP)	EST. BONUS: GROSS P/Y (GBP)	CONTRACT DETAILS: SIGNED	CONTRACT DETAILS: EXPIRATION	CONTRACT DETAILS: YEARS REMAINING	CONTRACT DETAILS: GROSS REMAINING (GBP)	CONTRACT DETAILS: RELEASE CLAUSE (GBP)	BIO: STATUS	BIO: POS.	BIO: POS.	BIO: AGE	BIO: COUNTRY
Search					Search								>,>=,<=,<	Search
Kevin De Bruyne	✓	£ 400,000	£ 20,800,000		Apr 7, 2021	Jun 30, 2025	2	£ 41,600,000		n/a	F	AM	32	Belgium
Erling Haaland	✓	£ 375,000	£ 19,500,000	£ 24,700,000	Jul 1, 2022	Jun 30, 2027	4	£ 78,000,000	£ 173,000,000	n/a	F	CF	23	Norway
Casemiro	✓	£ 350,000	£ 18,200,000	£ 5,200,000	Aug 22, 2022	Jun 30, 2026	3	£ 54,600,000		n/a	M	DM	31	Brazil
Mohamed Salah	✓	£ 350,000	£ 18,200,000		Jul 1, 2022	Jun 30, 2025	2	£ 36,400,000		n/a	F	RW	31	Egypt
Jadon Sancho	✓	£ 350,000	£ 18,200,000		Jul 23, 2021	Jun 30, 2026	3	£ 54,600,000		n/a	F	LW	23	England
Raphaël Varane	✓	£ 340,000	£ 17,680,000		Aug 14, 2021	Jun 30, 2025	2	£ 35,360,000		n/a	D	CB	30	France
Raheem Sterling	✓	£ 325,000	£ 16,900,000		Jul 13, 2022	Jun 30, 2027	4	£ 67,600,000		n/a	F	LW	28	England
Marcus Rashford	✓	£ 300,000	£ 15,600,000	£ 3,900,000	Jul 18, 2023	Jun 30, 2028	5	£ 78,000,000		n/a	F	LW	25	England
Bernardo Silva	✓	£ 300,000	£ 15,600,000		Aug 23, 2023	Jun 30, 2026	3	£ 46,800,000	£ 45,000,000	n/a	F	AM	29	Portugal
Jack Grealish	✓	£ 300,000	£ 15,600,000		Aug 5, 2021	Jun 30, 2027	4	£ 62,400,000		n/a	F	LW	27	England
Kai Havertz	✓	£ 280,000	£ 14,560,000	£ 2,600,000	Jul 1, 2023	Jun 30, 2028	5	£ 72,800,000		n/a	F	AM	24	Germany
Gabriel Jesus	✓	£ 265,000	£ 13,780,000		Jul 4, 2022	Jun 30, 2027	4	£ 55,120,000		n/a	F	CF	26	Brazil
Mason Mount	✓	£ 250,000	£ 13,000,000	£ 2,600,000	Jul 5, 2023	Jun 30, 2028	5	£ 65,000,000		n/a	F	AM	24	England
John Stones	✓	£ 250,000	£ 13,000,000		Aug 10, 2021	Jun 30, 2026	3	£ 39,000,000		n/a	D	CB	29	England
Anthony Martial	✓	£ 250,000	£ 13,000,000		Sep 1, 2015	Jun 30, 2024	1	£ 13,000,000		n/a	F	CF	27	France
Reece James	✓	£ 250,000	£ 13,000,000		Sep 5, 2022	Jun 30, 2028	5	£ 65,000,000		n/a	D	RB	23	England
Bruno Fernandes	✓	£ 240,000	£ 12,480,000		Jan 29, 2020	Jun 30, 2026	3	£ 37,440,000		n/a	F	AM	28	Portugal
Declan Rice	✓	£ 240,000	£ 12,480,000		Jul 15, 2023	Jun 30, 2028	5	£ 62,400,000		n/a	M	DM	24	England

```
Name,Weekly Salary,Base Salary,ADJ Salary,Team,year
Radamel Falcao,285000,14820000,17852178,Manchester United,2014
Angel Di Maria,250000,13000000,15659805,Manchester United,2014
Wayne Rooney,235000,12220000,14720217,Manchester United,2014
Robin van Persie,210000,10920000,13154236,Manchester United,201
```

❑ 사이트 정보

- ❑ 사이트명 : capology
- ❑ 사이트 특징 : 선수들의 2014~2022 연봉 제공
- ❑ 전체 데이터 수 : 6140
- ❑ 결측치 : -
- ❑ 사용피쳐 : Name, Weekly Salary, Base Salary, ADJ Salary
- ❑ 주의사항 :
  - ❑ 기본적으로 100명의 선수를 보여주지만 All 버튼을 통해 당해년도 모든 선수를 볼 수 있음
  - ❑ 크롬 개발자 도구에서는 정상적으로 데이터가 나오지만 크롤링시 중복 데이터가 발생
- ❑ 수집결과 : capology_(2014~2022).csv

결측률 확인(2)

2차 완성 된 데이터 통합 결과!


```
eda_df.isnull().mean()
```

	data
Id	0.00000
No.	0.00000
year	0.00000
Name	0.00000
Age	0.00000
Team	0.00000
Position	0.00000
Weekly Salary	0.06413
Base Salary	0.06413
ADJ.Gross Salary	0.06413

- ❑ 결과 내용
  - ❑ 기존 스탯 정보에 있는 선수들 매칭률 증가
  - ❑ **결측률 약 34퍼센트 -> 약 6퍼센트로 감소**
- ❑ 문제점 발견
  - ❑ 이름을 별명으로 사용하는 선수 발견
  - ❑ 한국선수들 성과 이름 반대로 된 상태 발견
  - ❑ 소수의 오타 발견
  - ❑ 중복인 이름 발견
- ❑ 해결 방안
  - ❑ 이름이 다른 선수들 outer join 으로 리스트를 만들어서 player 테이블을 기준으로 변환!

수작업을 통한 전처리


```
detail_list =  {'Ahmed Elmohamady': 'Ahmed El Mohamady',
                'Alex Oxlade-Chamberlain': 'Alex Oxlade Chamberlain',
                'Alex Song': 'Alexandre Song',
                'Robbie Brady': 'Robert Brady',
                'Cheik Tiote': 'Cheick Tiote',
                'Andrew Robertson': 'Andy Robertson',
                "Joey O'Brien": "Joseph O'Brien",
                'Rob Green': 'Robert Green',
                'Papiss Demba Cisse': 'Papiss Cisse',
                'Falcao': 'Radamel Falcao',
                'Matthew Upson': 'Matt Upson',
                'Jonathan Williams': 'Jonny Williams',
                'John Mikel Obi': 'John Obi Mikel',
                'Rob Elliot': 'Robert Elliot',
                'Will Buckley': 'William Buckley',
                'Brad Jones': 'Bradley Jones',
                'Tyias Browning': 'Jiang Guangtai',
```

이름이 조금씩 차이나는 선수들은 정렬된 리스트를 보며 하드코딩으로 매칭해서 변환

```
import pandas as pd

def no_convert(num,nam):
    # CSV파일 읽기
    csv_path = 'C:/Users/LEGION/Downloads/understat_all.csv'
    df = pd.read_csv(csv_path)

    for i in range(0,len(df[df['No'] == num])):
        # 'No'칼럼에서 원하는 번호 찾기
        row_index = df[df['No'] == num].index[i]

        # 새로운 값으로 바꿔주기
        new_name = nam
        df.at[row_index, 'Player'] = new_name
```

이름이 중복인 선수들은 고유넘버로 매칭해서 변경
(아스널이라는 팀에는 가브리엘이라는 선수가 4명 있었음)

결측률 확인(3)

최종 완성 된 데이터 통합 결과!

```
eda_df.isnull().mean()
```

```
Id              0.000000
No.             0.000000
year            0.000000
Name            0.000000
Age             0.000000
Team            0.000000
Position        0.000000
Weekly Salary   0.033221
Base Salary     0.033221
ADJ Salary      0.033221
dtype: float64
```

❑ 결과 내용

- ❑ 기존 스탯 정보에 있는 선수들과 매칭률 증가
- ❑ 누락된 결측치를 제외한 최대한의 매칭 성공
- ❑ **최종적으로 결측률이 약 3퍼센트로 감소**

데이터 요약

분석 대상

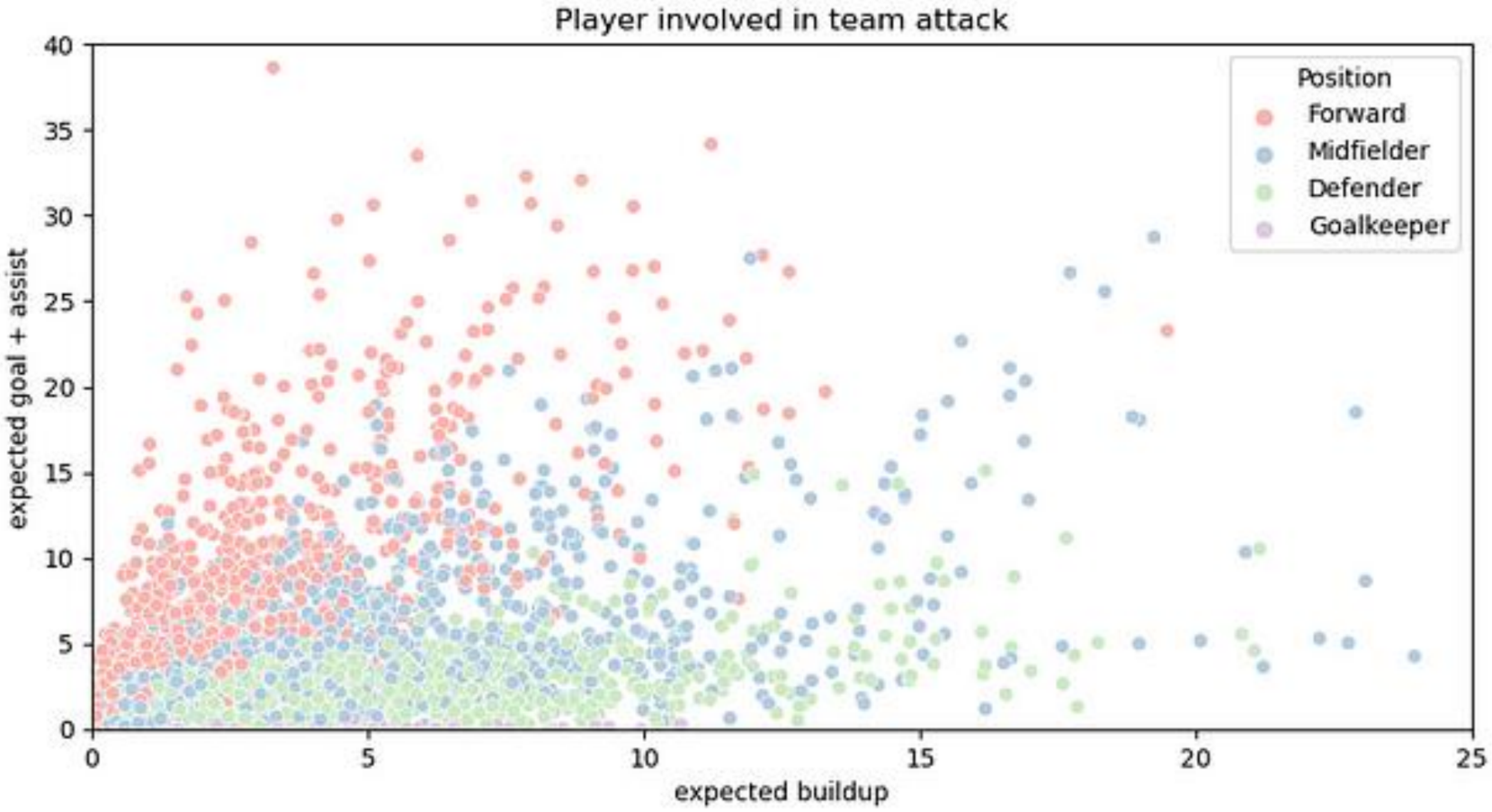
대상	EPL(Premier League)	분석 연도	2014~2022
피쳐수	45 개	총행	4588 행
선수	1641 명	팀수	33 팀

피처 분류

항목		개수	비고
수치형	연속형	32개	G, NPG, A, xG, NPxG, xA, xGChain, xGBuildup, xG90, NPxG90, xA90, xG90+xA90, NPxG90+xA90, xGChain90, xGBuildup90, SpG, KeyP, Drb_Off, Fouled, Off, Disp, UnsTch, Rating, Tackles, Inter, Fouls, Offsides, Clear, Drb_Def, Blocks, AvgP, PS%
	이산형	8개	Birth Year, Age, year, Weekly Salary, Base Salary, ADJ Salary, Apps, Min
범주형	순서형	1개	Age Lev
	명목형	4개	Player Id, Name, Team, Position

content

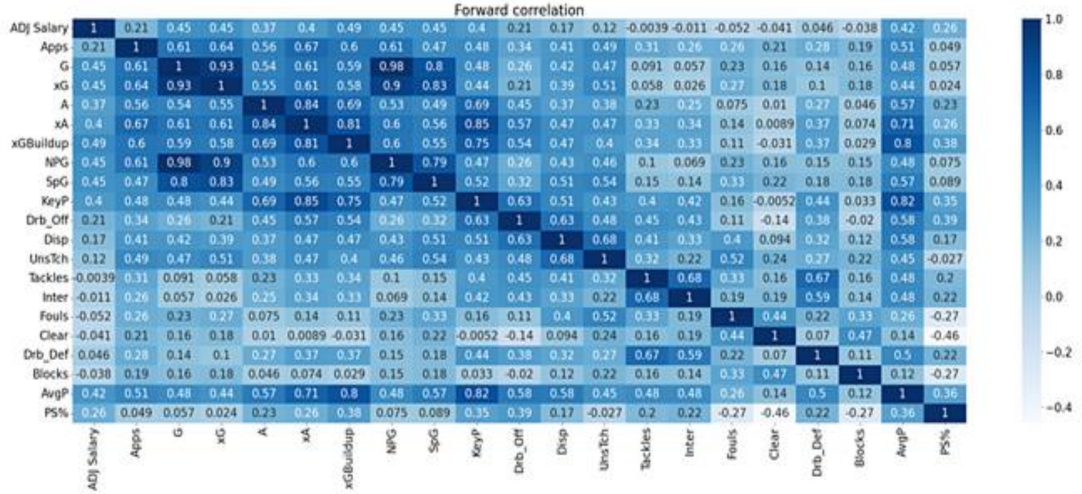
포지션별 주요 스텟 분석

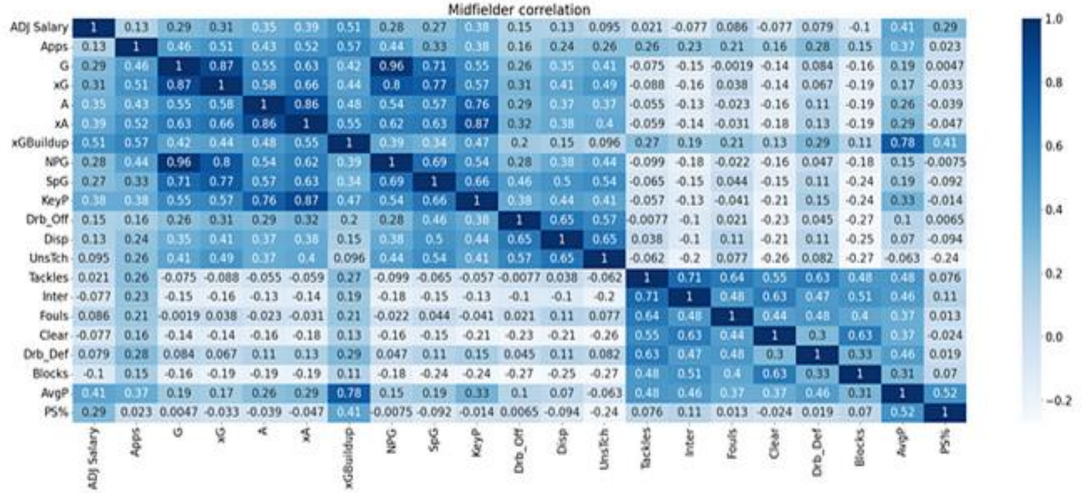


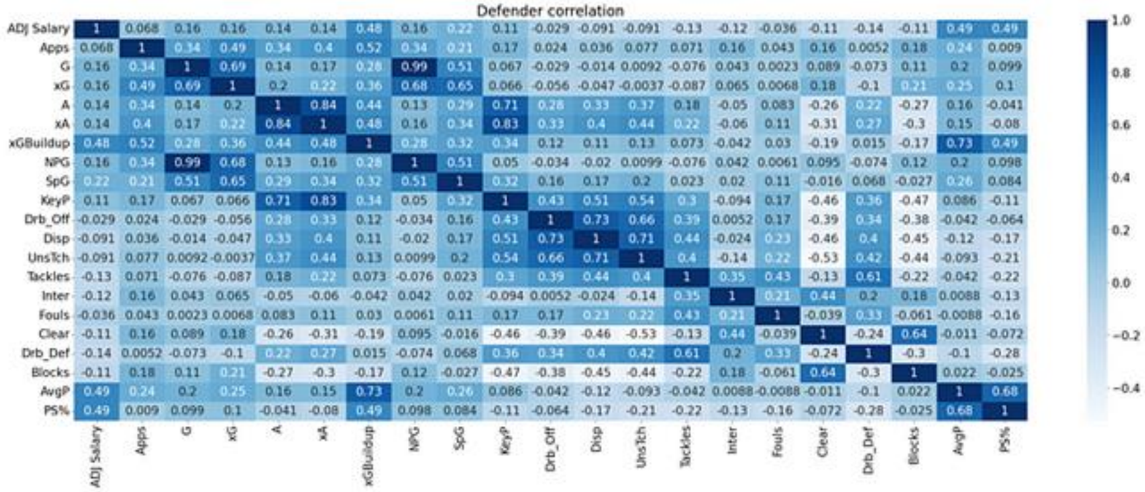
공격수 : 높은 xG + xA 수치 → **공격의 마무리에 많이 가담**

수비수와 미드필더 : 높은 xGBuildup 수치 → **공격 과정에 많이 가담**

포지션별 상관계수 분석







포지션 별로 **연봉과 상관관계가 있는 지표**는 어떤 것이 있는지 살펴보기 위해 상관계수를 히트맵으로 표현

상관관계가 0.35 이상 피쳐 추출

Forward : G, xG, A,xA, xG Buildup, NPG, SpG, KeyP, AvgP

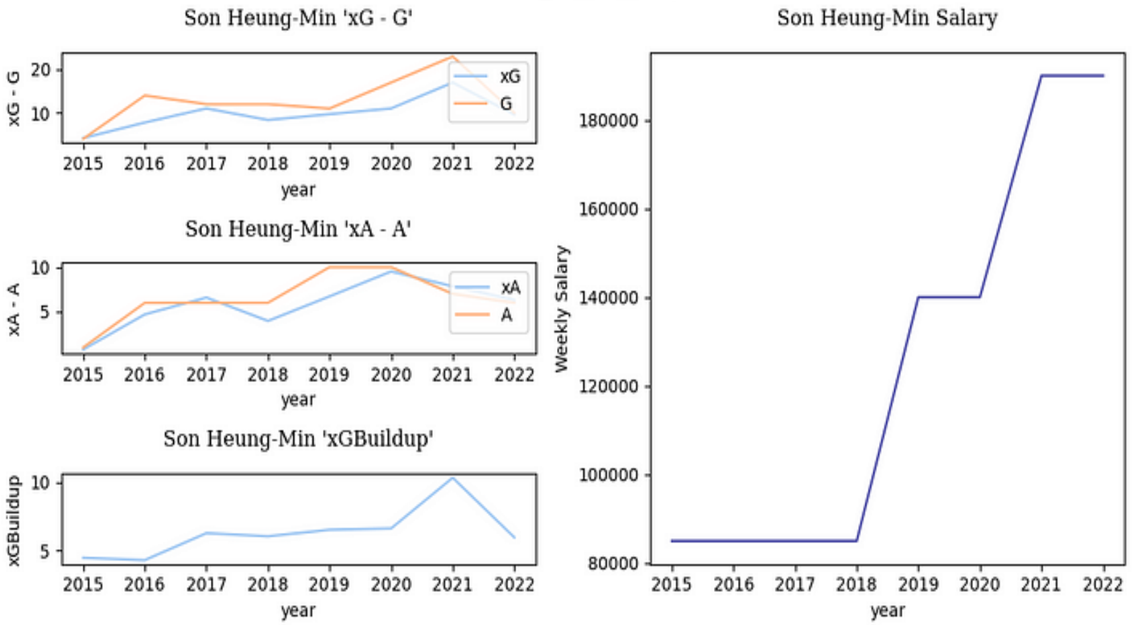
Midfielder : A, xA, xG Buildup, KeyP, AvgP

Defender : xG Buildup, AvgP, PS%

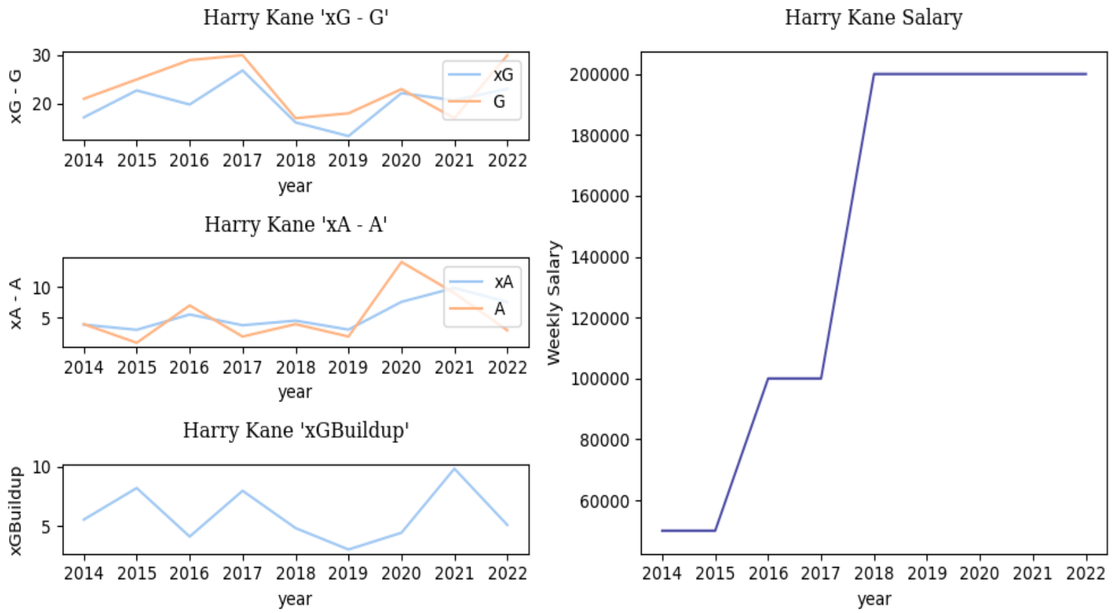
선수 분석 - 공격수

공격수의 주요 지표 중 xG, G, xA, A, xGBuildup과 연봉의 관계

손흥민 분석



해리 케인 분석



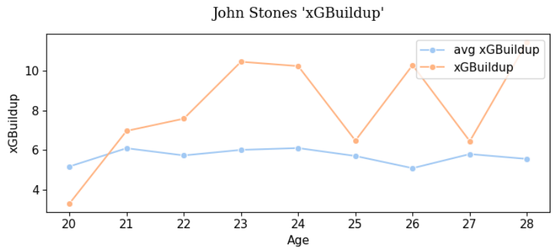
손흥민 선수와 **해리 케인** 선수가
기대 수치 보다 높은 득점 및 도움을 기록을 했을 때,

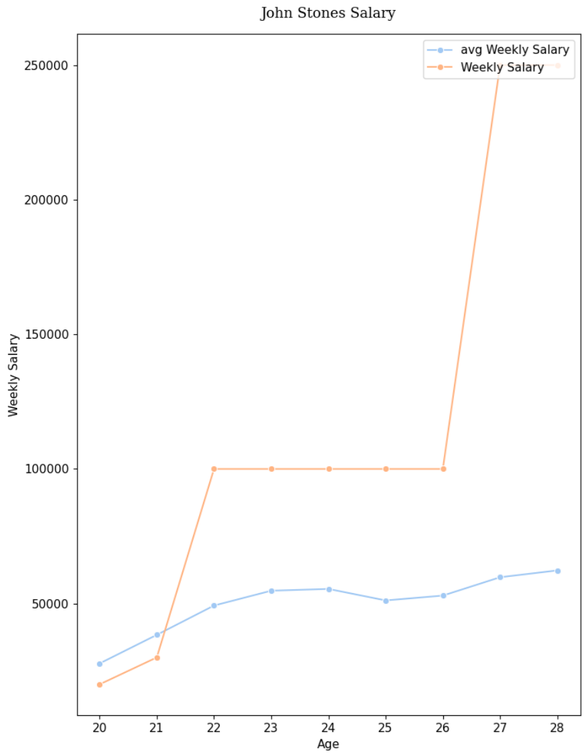
즉 **좋은 기량을 보인 이후**에 **연봉이 상승**하는 것을 확인 할 수 있음

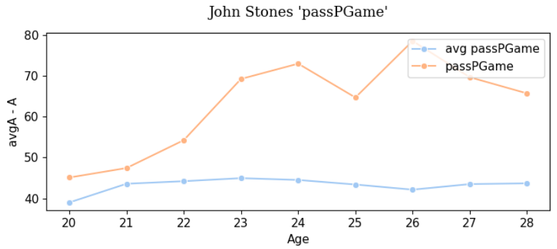
선수 분석 - 수비수

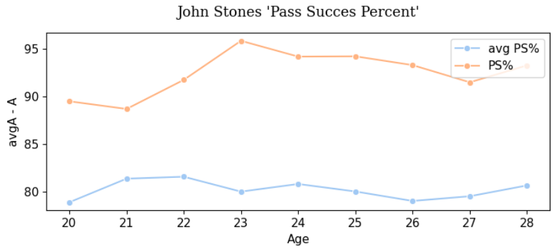
수비수의 주요 지표 중 xGBuildup, AvgP, PS% 와 연봉의 관계

존 스톤스 분석

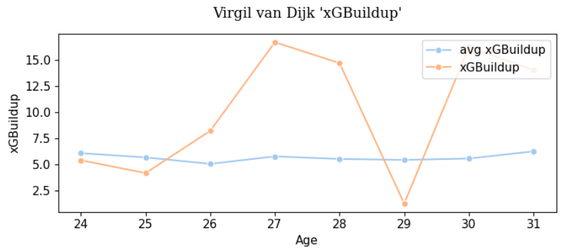


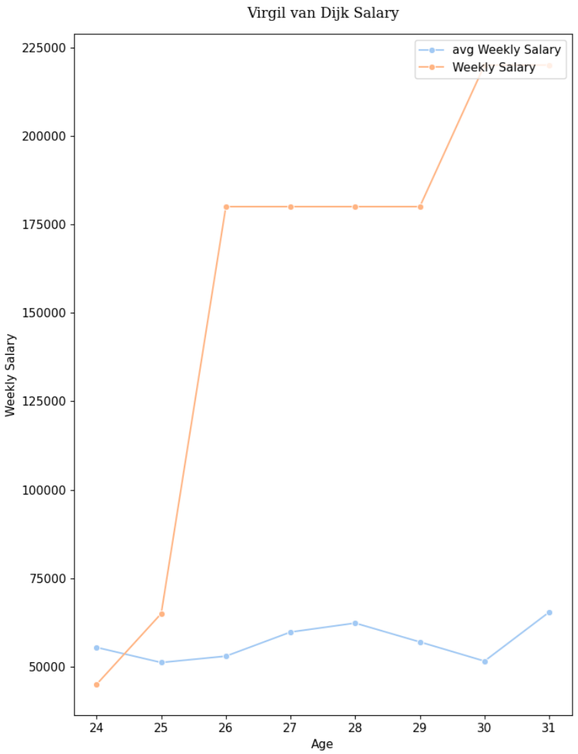


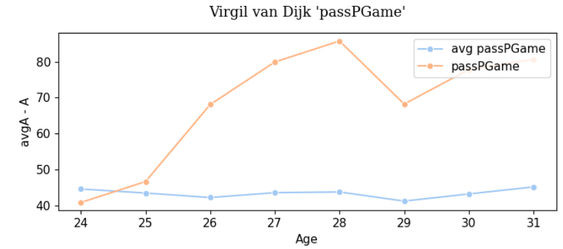


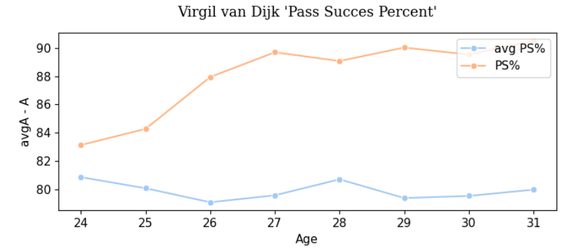


반 다이크 분석









존 스톤스 선수와 **반 다이크** 선수가
동 나이, 동 포지션 보다 좋은 스탯을 기록을 했을 때,

즉 **좋은 기량을 보인 이후**에 **연봉이 상승**하는 것을 확인 할 수 있음

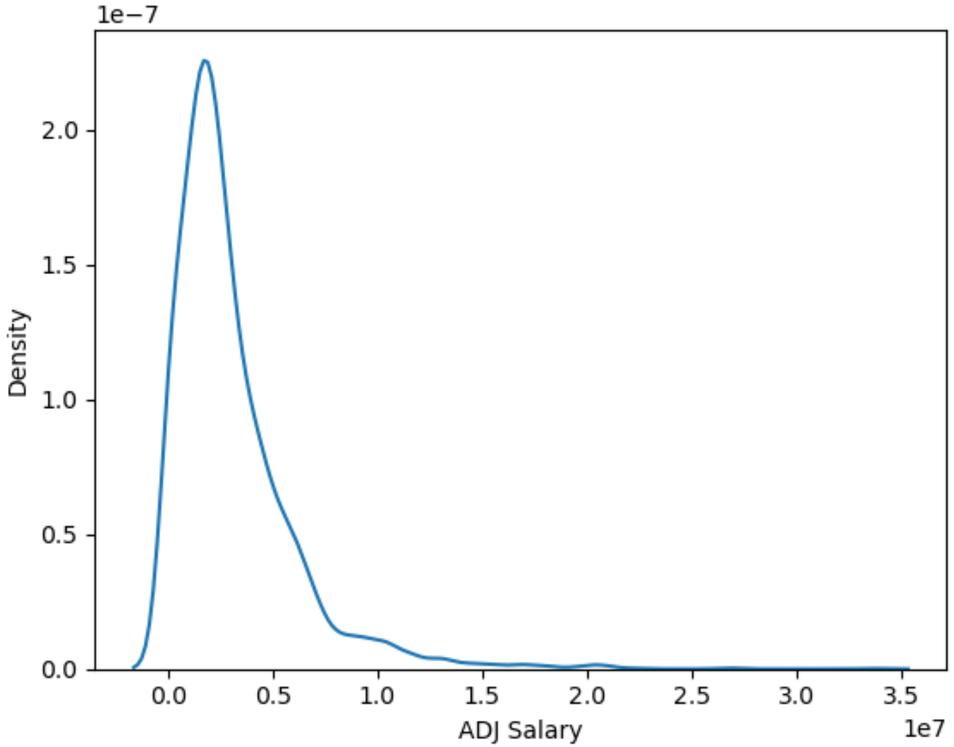
유사도 분석 - 스케일 변환

Feature의 스케일 변환

- PDF 형상을 유지하기 위해 RobustScaler 이용

```
sns.kdeplot(eda_df, x='ADJ Salary')
```

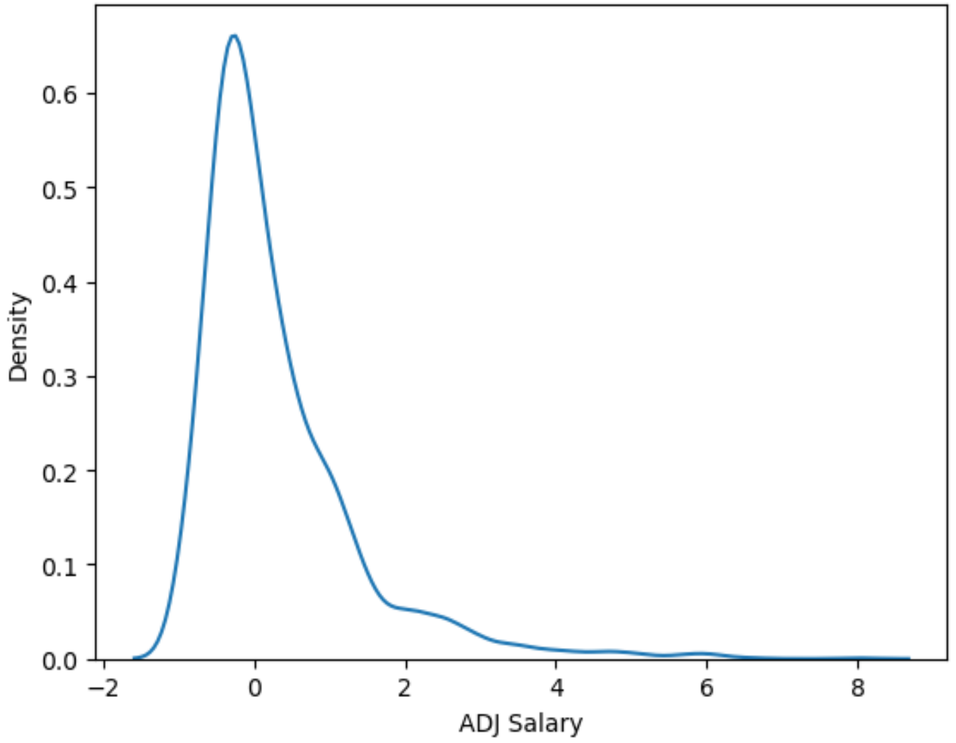
```
<Axes: xlabel='ADJ Salary', ylabel='Density'>
```



scale 변환 전 Density Curve

```
sns.kdeplot(scalered_df, x='ADJ Salary')
```

```
<Axes: xlabel='ADJ Salary', ylabel='Density'>
```



scale 변환 후 Density Curve

유사도 분석 - consine similarity

유사도 분석

- 손흥민 선수의 활약한 연도별 스탯 유사도 도출
- **cosine similarity** 이용

```
player_name = 'Son Heung-Min'
player_indexes = featuring_df[featuring_df['Name']==player_name].index
similarity = cosine_similarity(scalered_df.loc[list(player_indexes)], scalered_df)
similarity.T
```

유사도 분석결과 활용

- 도출된 연도별 스탯 유사도를 합하여 **가장 높은 값**을 가진 선수 선정

```
featuring_df['similarity'] = similarity.sum(axis=0)
featuring_df['similarity-rank'] = featuring_df['similarity'].rank(ascending=False)
featuring_df = featuring_df[featuring_df['Name']!=player_name].sort_values('similarity-rank')
featuring_df[~featuring_df.duplicated(['Name'], keep='first')].iloc[:3, 1:7]
```

- 유사도 분석 결과

	0	1	2	3	4	5	6	7
0	0.395899	0.700208	0.730967	0.690117	0.801017	0.801899	0.790609	0.753354
1	0.537487	0.832247	0.845352	0.820112	0.910729	0.898351	0.896476	0.854849
2	0.741420	0.938697	0.953791	0.951678	0.916432	0.876917	0.945695	0.919144
3	0.154271	0.418792	0.466136	0.429066	0.626162	0.639044	0.510116	0.529246
4	0.331232	0.601542	0.679825	0.623773	0.771201	0.761318	0.709994	0.688781
5	0.455456	0.674785	0.697879	0.687370	0.760319	0.710790	0.675668	0.706308
6	-0.180782	0.005747	-0.036945	-0.006744	0.075414	0.113018	0.094753	-0.001808
7	0.459840	0.707495	0.762430	0.707130	0.820247	0.850041	0.832929	0.781338
8	-0.392677	-0.307104	-0.333354	-0.308877	-0.269724	-0.279978	-0.261307	-0.359585
9	-0.024001	0.126057	0.131079	0.159567	0.224342	0.217280	0.236560	0.081332

- 활용 예시

	Name	Birth Year	Age	Team	Position	year
1666	Alexandre Lacazette	1991	26	Arsenal	Forward	2017
1606	Sergio Aguero	1988	29	Man City	Forward	2017
2122	Mohamed Salah	1992	26	Liverpool	Forward	2018

유사도 분석 - euclidean distance

유사도 분석

- 손흥민 선수의 활약한 연도별 스탯 유사도 도출
- **euclidean distance** 이용

```
player_name = 'Son Heung-Min'
player_indexes = featuring_df[featuring_df['Name']==player_name].index
similarity = euclidean_distances(scalered_df.loc[list(player_indexes)], scalered_df)
similarity.T
```

유사도 분석결과 활용

- 도출된 연도별 스탯 유사도를 합하여 **가장 낮은 값**을 가진 선수 선정

```
featuring_df['similarity'] = similarity.sum(axis=0)
featuring_df['similarity-rank'] = featuring_df['similarity'].rank(ascending=True)
featuring_df = featuring_df[featuring_df['Name']!=player_name].sort_values('similarity-rank')
featuring_df[~featuring_df.duplicated(['Name'], keep='first')].iloc[:3, 1:7]
```

- 유사도 분석 결과

	0	1	2	3	4	5	6	7
0	12.390015	9.508504	8.989546	9.640049	8.131633	7.928352	8.635090	8.974820
1	10.848308	7.309165	6.913749	7.524186	5.710746	5.616972	6.205897	6.996994
2	13.927301	9.165861	8.257785	9.101656	8.993657	8.908588	5.907172	9.669901
3	10.343784	9.877270	9.805380	9.752845	8.194911	8.420485	11.890897	8.872108
4	9.202360	8.092547	7.595362	7.899777	6.262465	6.760314	9.472447	6.865858
5	6.781681	6.198909	6.498665	6.066030	5.866982	6.975086	9.914385	5.618805
6	7.163579	9.453360	10.261521	9.486005	9.705878	10.403012	13.311480	9.053930
7	9.507946	7.418618	6.971429	7.527241	6.053075	5.369815	7.425014	6.433785
8	7.851825	10.725065	11.406921	10.638060	11.162701	12.136127	14.899277	10.430538
9	6.385332	8.696693	9.277679	8.494997	8.762163	9.681890	12.485171	8.260645

- 활용 예시

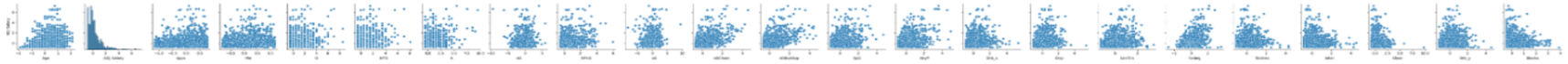
	Name	Birth Year	Age	Team	Position	year
3678	Phil Foden	2000	21	Man City	Midfielder	2021
2212	Marcus Rashford	1998	20	Man Utd	Forward	2018
1666	Alexandre Lacazette	1991	26	Arsenal	Forward	2017

content

데모 페이지 개발 동기


```
sns.pairplot(mid_df, x_vars=mid_df.columns, y_vars=['ADJ Salary'])
```

```
<seaborn.axisgrid.PairGrid at 0x2d96ea33cd0>
```

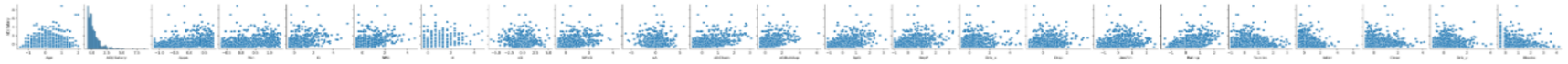


```
sns.pairplot(def_df, x_vars=def_df.columns, y_vars=['ADJ Salary'])
```

```
<seaborn.axisgrid.PairGrid at 0x2d96ea33d60>
```

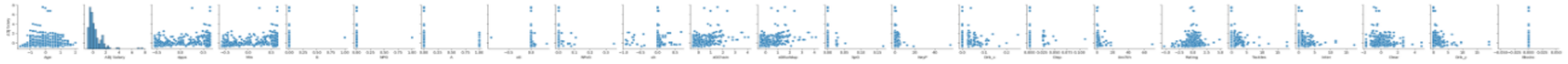
```
sns.pairplot(for_df, x_vars=for_df.columns, y_vars=['ADJ Salary'])
```

```
<seaborn.axisgrid.PairGrid at 0x2d970709270>
```



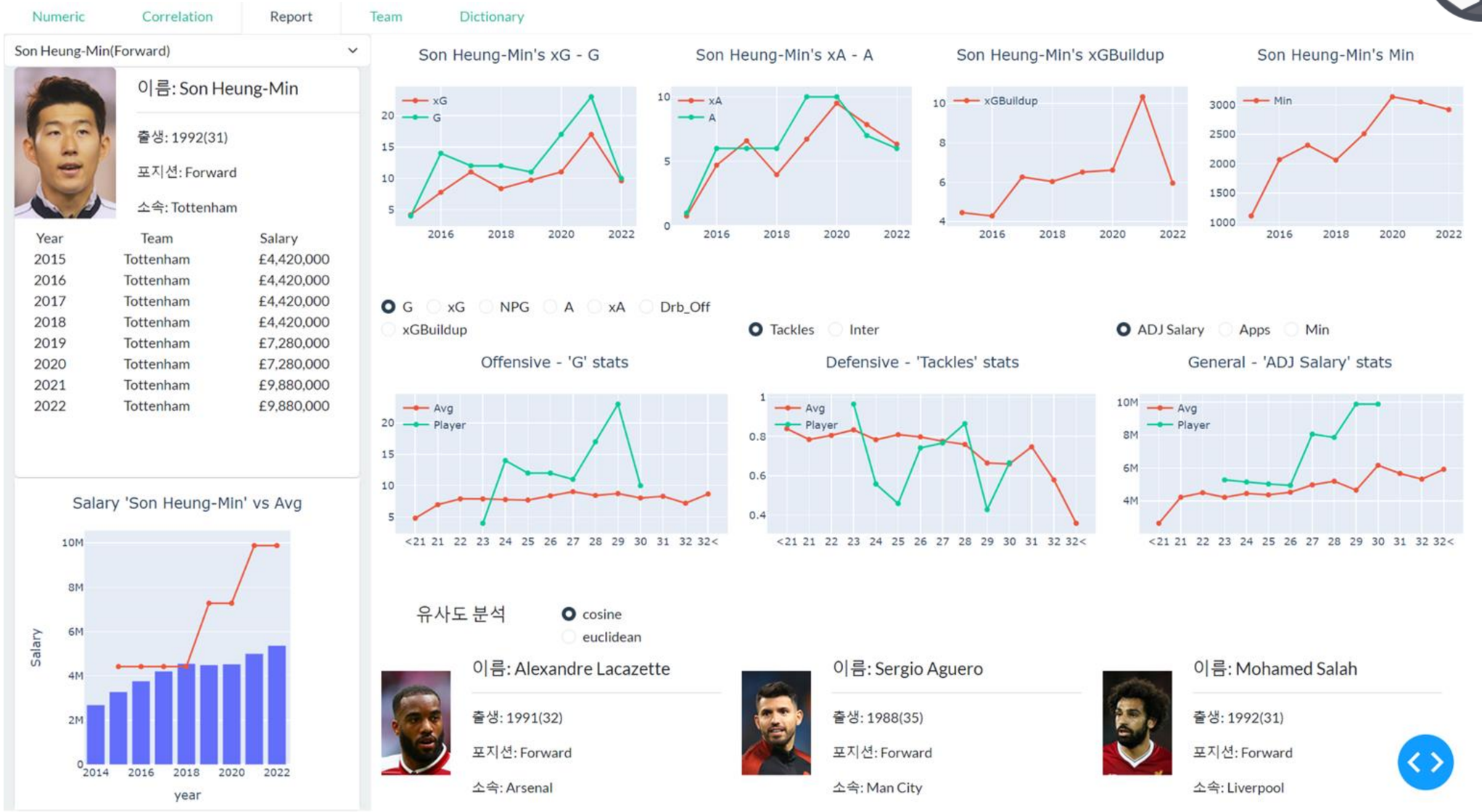
```
sns.pairplot(keep_df, x_vars=keep_df.columns, y_vars=['ADJ Salary'])
```

```
<seaborn.axisgrid.PairGrid at 0x2d9732bab60>
```



40개가 넘는 피처를 쉽게 분석하기 위해서 웹 페이지 만들기 시작

데모 페이지 시연


Numeric
Correlation
Report
Team
Dictionary
Son Heung-Min(Forward)
이름: Son Heung-Min
출생: 1992(31)
포지션: Forward
소속: Tottenham
Year Team Salary
2015 Tottenham £4,420,000
2016 Tottenham £4,420,000
2017 Tottenham £4,420,000
2018 Tottenham £4,420,000
2019 Tottenham £7,280,000
2020 Tottenham £7,280,000
2021 Tottenham £9,880,000
2022 Tottenham £9,880,000
Salary 'Son Heung-Min' vs Avg
Salary
year
Son Heung-Min's xG - G
Son Heung-Min's xA - A
Son Heung-Min's xGBuildup
Son Heung-Min's Min
G xG NPG A xA Drb_Off
xGBuildup
Tackles Inter
ADJ Salary Apps Min
Offensive - 'G' stats
Defensive - 'Tackles' stats
General - 'ADJ Salary' stats
Avg
Player
유사도 분석
cosine
euclidean
이름: Alexandre Lacazette
출생: 1991(32)
포지션: Forward
소속: Arsenal
이름: Sergio Aguero
출생: 1988(35)
포지션: Forward
소속: Man City
이름: Mohamed Salah
출생: 1992(31)
포지션: Forward
소속: Liverpool

content

참고 사이트

- EPL 선수들의 수비적인 것과 공격적인 세부 스탯 정보

https://1xbet.whoscored.com/Regions/252/Tournaments/2/England-Premier-League

- EPL 선수들의 시즌별 세부 스탯 정보

https://understat.com/league/EPL

- EPL 선수들의 연봉정보

https://www.capology.com/uk/premier-league/salaries

- github(김동억)

https://github.com/sajacaros/eda_champions

- github(이도형)

https://github.com/Lee-Dohyeong/EDA_Project_

- github(이인서)

https://github.com/LEEINSEO-0118/EPL_EDA_Project

- github(권도형)

https://github.com/dhkwon1984/EDA__Champions/tree/main

content

7. Q&A

Q&A

발표를 봐주셔서 감사합니다.

자유롭게 질문해 주세요.