

スクレイピングで NBA選手のデータセット作成

2-3-30 佐藤佑樹

- ▶ NBAが好きでよく見るから
- ▶ 色々数字があるけど、どの数字が重要なのか分かりにくい

Regular Season		Playoffs																															
Rk	Player	Age	Team	Pos	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	2P	2PA	2P%	eFG%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS	Awards			
1	Joel Embiid	29	PHI	C	39	39	33.6	11.5	21.8	.529	1.4	3.6	.388	10.2	18.3	.556	.561	10.2	11.6	.883	2.4	8.6	11.0	5.6	1.2	1.7	3.8	2.9	34.7	AS			
2	Luka Dončić	24	DAL	PG	70	70	37.5	11.5	23.6	.487	4.1	10.6	.382	7.4	13.0	.573	.573	6.8	8.7	.786	0.8	8.4	9.2	9.8	1.4	0.5	4.0	2.1	33.9	MVP-3,CPOY-6,AS,NBA1			
3	Giannis Antetokounmpo	29	MIL	PF	73	73	35.2	11.5	18.8	.611	0.5	1.7	.274	11.0	17.1	.645	.624	7.0	10.7	.657	2.7	8.8	11.5	6.5	1.2	1.1	3.4	2.9	30.4	MVP-4,DPOY-9,CPOY-12,AS,NBA2			
4	Shai Gilgeous-Alexander	25	OKC	PG	75	75	34.0	10.6	19.8	.535	1.3	3.6	.353	9.3	16.2	.576	.567	7.6	8.7	.874	0.9	4.7	5.5	6.2	2.0	0.9	2.2	2.5	30.1	MVP-2,DPOY-7,CPOY-3,AS,NBA1			
5	Jalen Brunson	27	NYK	PG	77	77	35.4	10.3	21.4	.479	2.7	6.8	.401	7.5	14.6	.516	.543	5.5	6.5	.847	0.6	3.1	3.6	6.7	0.9	0.2	2.4	1.9	28.7	MVP-5,CPOY-5,AS,NBA2			
6	Devin Booker	27	PHO	PG	68	68	36.0	9.4	19.2	.492	2.2	6.1	.364	7.2	13.0	.552	.550	6.0	6.7	.886	0.8	3.7	4.5	6.9	0.9	0.4	2.6	3.0	27.1	AS,NBA3			
7	Kevin Durant	35	PHO	PF	75	75	37.2	10.0	19.1	.523	2.2	5.4	.413	7.8	13.7	.567	.581	4.8	5.6	.856	0.5	6.1	6.6	5.0	0.9	1.2	3.3	1.8	27.1	MVP-9,AS,NBA2			
8	Jayson Tatum	25	BOS	PF	74	74	35.7	9.1	19.3	.471	3.1	8.2	.376	6.0	11.0	.542	.552	5.6	6.7	.833	0.9	7.2	8.1	4.9	1.0	0.6	2.5	2.0	26.9	MVP-6,CPOY-9,AS,NBA1			
9	De'Aaron Fox	26	SAC	PG	74	74	35.9	9.7	20.9	.465	2.9	7.8	.369	6.8	13.1	.522	.534	4.2	5.7	.738	0.9	3.7	4.6	5.6	2.0	0.4	2.6	2.6	26.6	CPOY-12			
10	Donovan Mitchell	27	CLE	SG	55	55	35.3	9.1	19.8	.462	3.3	9.0	.368	5.8	10.8	.539	.545	5.0	5.8	.865	0.8	4.3	5.1	6.1	1.8	0.5	2.8	2.1	26.6	AS			
11	Stephen Curry	35	GSW	PG	74	74	32.7	8.8	19.5	.450	4.8	11.8	.408	4.0	7.7	.515	.573	4.0	4.4	.923	0.5	4.0	4.5	5.1	0.7	0.4	2.8	1.6	26.4	CPOY-1,AS,NBA3			
12	Nikola Jokić	28	DEN	C	79	79	34.6	10.4	17.9	.583	1.1	2.9	.359	9.4	14.9	.626	.612	4.5	5.5	.817	2.8	9.5	12.4	9.0	1.4	0.9	3.0	2.5	26.4	MVP-1,CPOY-4,AS,NBA1			
13	Anthony Edwards	22	MIN	SG	79	78	35.1	9.1	19.7	.461	2.4	6.7	.357	6.7	13.0	.515	.522	5.4	6.4	.836	0.7	4.8	5.4	5.1	1.3	0.5	3.1	1.8	25.9	MVP-7,CPOY-8,AS,NBA2			
14	Tyrese Maxey	23	PHI	PG	70	70	37.5	9.1	20.3	.450	3.0	8.1	.373	6.1	12.1	.501	.524	4.7	5.4	.868	0.5	3.2	3.7	6.2	1.0	0.5	1.7	2.2	25.9	MIP-1,AS			
15	LeBron James	39	LAL	PF																													

今回使用するデータ

- ▶ URL: https://www.basketball-reference.com/leagues/NBA_2024_per_game.html
- ▶ 2023-2024シーズンのレギュラーシーズンの
1 試合平均スタッツ

スクレイピングしたデータ

- ▶ 選手名
- ▶ 歳
- ▶ 所属チーム
- ▶ FG% : 試合中に放たれる 2 ポイントシュートと 3 ポイントシュートの入る確率
- ▶ PPG(Points Per Game) : 1試合あたりの平均得点

実際に使ったコード

```
1  # ライブラリーのインポート
2  import requests
3  from bs4 import BeautifulSoup
4  import pandas as pd
5
6  # URLにアクセスし、HTMLを取得
7  url = "https://www.basketball-reference.com/leagues/NBA_2024_per_game.html"
8  response = requests.get(url, headers=headers)
9  soup = BeautifulSoup(response.text, "html.parser")
10 # per_game_stats というIDのtableを取得
11 table = soup.find("table", id="per_game_stats")
12
13 # tableデータの抽出
14 tbody = table.find("tbody")
15 rows = tbody.find_all("tr")
16 players_data = []
17
18 for row in rows:
19     # 行の中に <td data-stat="name_display"> が無ければスキップ
20     name_cell = row.find("td", {"data-stat": "name_display"})
21     if not name_cell:
22         continue
23
24     name = name_cell.get_text(strip=True)
```

HTMLの構造

```
▼<table class="stats_table sortable soc now_sortable is_sorted" id="per_game_stats"
  data-cols-to-freeze="2,4" data-non-qual="1" data-qual-text="" data-qual-label="When t
  able is sorted, hide non-qualifiers for rate stats" data-soc-sum="" data-type="player
  _season" data-soc-sum-phase-type="reg">
  <caption>Per Game Table</caption>
  ▶<colgroup>⋮</colgroup>
  ▼<thead>
  ▶<tr>⋮</tr>
  </thead>
  ▼<tbody>
  ▼<tr data-row="1">
    <th scope="row" class="right" data-stat="ranker" data-cs="2" data-kind="parent">1</th>
    ▼<td class="left" data-append-csv="doncilu01" data-stat="name_display" csk="
      "Dončić Luka-1">== $0
      <a href="/players/d/doncilu01.html">Luka Dončić</a>
    </td>
    <td class="right" data-stat="age">24</td>
    ▶<td class="left" data-stat="team_name" data-cs="2" data-kind="parent">⋮</td>
    <td class="center" data-stat="pos">PG</td>
    <td class="right" data-stat="games">70</td>
    <td class="right" data-stat="games_started">70</td>
    <td class="right" data-stat="mp_per_g" csk="37.4857142857">37.5</td>
    <td class="right" data-stat="fg_per_g" csk="11.4857142857">11.5</td>
    <td class="right" data-stat="fga_per_g" csk="23.6000000000">23.6</td>
    <td class="right" data-stat="fg_pct" csk="0.4866828087">.487</td>
    <td class="right" data-stat="fg3_per_g" csk="4.0571428571">4.1</td>
    <td class="right" data-stat="fg3a_per_g" csk="10.6285714286">10.6</td>
    <td class="right" data-stat="fg3_pct" csk="0.3817204301">.382</td>
```

実際に使ったコード

```
25     # Age
26     age_cell = row.find("td", {"data-stat": "age"})
27     age = age_cell.get_text(strip=True) if age_cell else ""
28     # Team
29     team_cell = row.find("td", {"data-stat": "team_name_abbr"})
30     team = team_cell.get_text(strip=True) if team_cell else ""
31     # FG%
32     fg_cell = row.find("td", {"data-stat": "fg_pct"})
33     fg_pct = fg_cell.get_text(strip=True) if fg_cell else ""
34     # PPG
35     pts_cell = row.find("td", {"data-stat": "pts_per_g"})
36     pts = pts_cell.get_text(strip=True) if pts_cell else ""
37
38     players_data.append({
39         "Name": name,
40         "Age": age,
41         "Team": team,
42         "FG%": fg_pct,
43         "PPG": pts,
44     })
45
46     # pandasでDataFrame化
47     df = pd.DataFrame(players_data)
48     print("選手数:", len(df))
49
50     # CSVに保存
51     df.to_csv("nba_players.csv", index=False, encoding="utf-8")
```

スクレイピングをした結果

nba_players

- ▶ NBAでは入れ替わりが激しいため736人

Name	Age	Team	FG%	PPG
Joel Embiid	29	PHI	0.529	34.7
Luka Dončić	24	DAL	0.487	33.9
Giannis Antetokounmpo	29	MIL	0.611	30.4
Shai Gilgeous-Alexander	25	OKC	0.535	30.1
Jalen Brunson	27	NYK	0.479	28.7
Devin Booker	27	PHO	0.492	27.1
Kevin Durant	35	PHO	0.523	27.1
Jayson Tatum	25	BOS	0.471	26.9
De'Aaron Fox	26	SAC	0.465	26.6
Donovan Mitchell	27	CLE	0.462	26.6
Stephen Curry	35	GSW	0.450	26.4
Nikola Jokić	28	DEN	0.583	26.4
Anthony Edwards	22	MIN	0.461	25.9
Tyrese Maxey	23	PHI	0.450	25.9
LeBron James	39	LAL	0.540	25.7