

Đồ án khoa học

— dữ liệu —

Trần Đại Chí - 18127070

Trần Minh Quang - 18127192

Nội dung



Giới thiệu đề tài

Thu thập dữ liệu

Khám phá dữ liệu

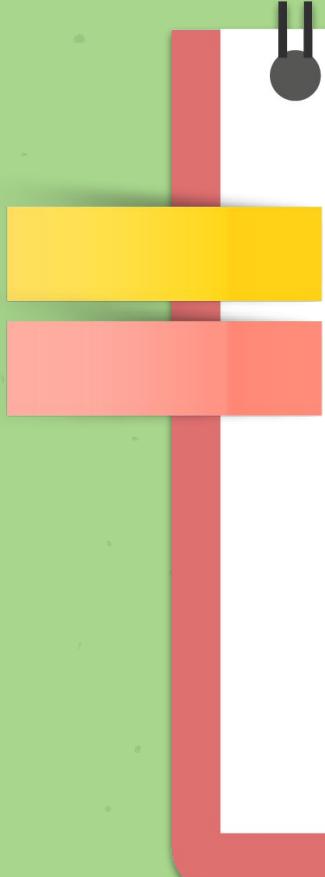


Trực quan hóa dữ liệu

Tiền xử lý dữ liệu

Mô hình hóa dữ liệu

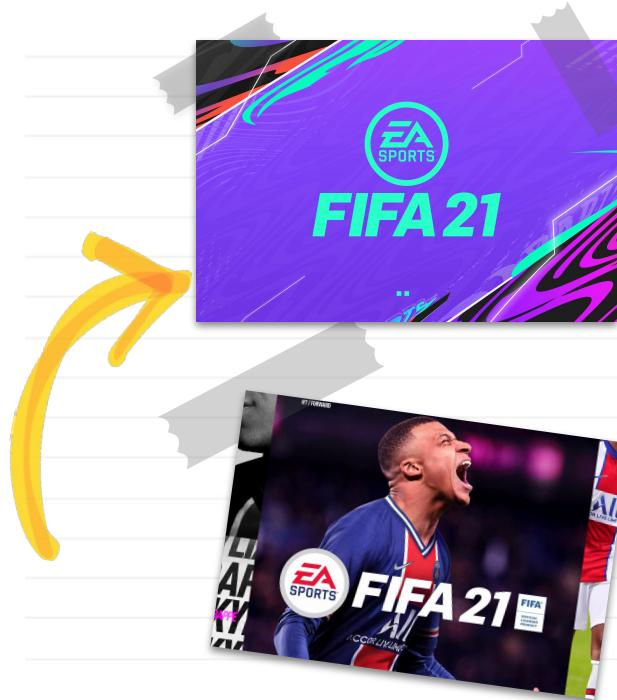


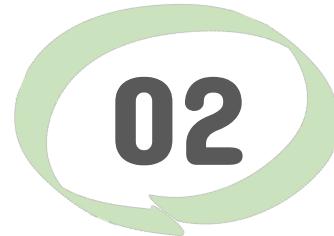


Giới thiệu đề tài

Giới thiệu đề tài

- FIFA 21 là một trò chơi mô phỏng bóng đá nổi tiếng được phát hành bởi Electronic Arts như là một phần của loạt game FIFA
- Vì sự hấp dẫn của tựa game này, ở đồ án này, nhóm sẽ thu thập dữ liệu về các cầu thủ bóng đá





Thu thập dữ liệu

Trang Web dùng để thu thập dữ liệu

sofifa.com/players?col=oa&sort=desc&offset=0

PLAYERS TEAMS SQUADS SHORTLISTS DISCUSSIONS SIGN IN ⚙️ 🇺🇸

Players
FIFA 21 ▾ SEP 2, 2021 ▾

All Added Updated Free On Loan Removed Customized Create Player Calculator Q Search Player ...

COLUMNS SELECTED

Age X Overall Rating X Potential X Value X Wage X Total Stats X Add Column

APPLY RESET

BASKET

COMPARE CLEAR + SHORTLIST + SQUAD

SEARCH

Name

All Players

Continents

Nationality / Region

Leagues

Teams

Age 15 45

NAME	AGE	JOVA	POT	TEAM & CONTRACT	VALUE	WAGE	TOTAL ...	HITS
L. Messi	33	93	93	Paris Saint-Germain 2021 ~ 2023	€103.5M	€320K	2231	400
R. Lewandowski	31	92	92	FC Bayern München 2014 ~ 2023	€124.5M	€270K	2211	207
Cristiano Ronaldo	35	92	92	Juventus 2018 ~ 2022	€63M	€220K	2221	373
J. Oblak	27	91	93	Atlético Madrid 2014 ~ 2023	€120M	€125K	1413	82
K. De Bruyne	29	91	91	Manchester City 2015 ~ 2025	€127.5M	€370K	2307	129
Neymar Jr	28	91	91	Paris Saint-Germain 2017 ~ 2025	€132M	€270K	2174	161
K. Mbappé	21	90	95	Paris Saint-Germain 2018 ~ 2022	€185.5M	€160K	2157	280
M. Salah	28	90	90	Liverpool 2017 ~ 2023	€120.5M	€250K	2213	97
V. van Dijk	28	90	91	Liverpool	€113M	€210K	2103	117

Thư viện dùng để thu thập dữ liệu



Scrapy

Bước 1: Thu thập ID của các cầu thủ

sc&offset=0

LISTS DISCUSSIONS SIGN IN 🚧 🇺🇸

removed Customized Create Player Calculator Q Search Player ...

NAME	AGE	IOVA	POT	TEAM & CONTRACT	VALUE	WAGE	TOTAL ...	HITS
L. Messi	33	93	93	Paris Saint-Germain 2021 ~ 2023	€103.5M	€320K	2231	400
R. Lewandowski	31	92	92	FC Bayern München 2014 ~ 2023	€124.5M	€270K	2211	207
Cristiano Ronaldo	35	92	92	Internazionale	€80M	€220K	2221	272

Elements Console Sources Network 1

```
<tbody class="list">
  <tr> = $0
    <td class="col-avatar" data-balloon-visible="true" data-balloon-pos="up">
      <figure class="avatar"></figure>
    </td>
    <td class="col-name">
      <a class="tooltip" href="/player/158023/lionel-messi" data-tooltip="Lionel Messi"></a>
        
      </td>
      <td class="col-ae" data-col="ae">33</td>
    <td class="col-ao col-sort" data-col="ao">93</td>
    <td class="col-pt" data-col="pt"></td>
    <td class="col-name"></td>
```

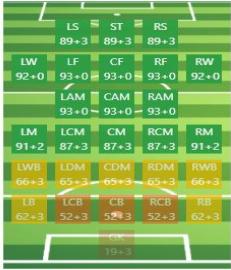
Bước 2: Lưu các ID đã thu thập được vào một file json và xóa các ID bị trùng trong file json này

```
2021-08-24 07:01:54 [scrapy.core.scrape] DEBUG: Scraped from <200 https://sofifa.com/players?col=oa&sort=desc&offset=16500>
{'player_url': '/player/200263'}
2021-08-24 07:01:54 [scrapy.core.scrape] DEBUG: Scraped from <200 https://sofifa.com/players?col=oa&sort=desc&offset=16500>
{'player_url': '/player/260935'}
2021-08-24 07:01:54 [scrapy.core.scrape] DEBUG: Scraped from <200 https://sofifa.com/players?col=oa&sort=desc&offset=16500>
{'player_url': '/player/239689'}
2021-08-24 07:01:54 [scrapy.core.scrape] DEBUG: Scraped from <200 https://sofifa.com/players?col=oa&sort=desc&offset=16500>
{'player_url': '/player/246345'}
2021-08-24 07:01:54 [scrapy.core.scrape] DEBUG: Scraped from <200 https://sofifa.com/players?col=oa&sort=desc&offset=16500>
{'player_url': '/player/257353'}
2021-08-24 07:01:54 [scrapy.core.scrape] DEBUG: Scraped from <200 https://sofifa.com/players?col=oa&sort=desc&offset=16500>
{'player_url': '/player/261449'}
2021-08-24 07:01:54 [scrapy.core.scrape] DEBUG: Scraped from <200 https://sofifa.com/players?col=oa&sort=desc&offset=16500>
{'player_url': '/player/261705'}
2021-08-24 07:01:54 [scrapy.core.scrape] DEBUG: Scraped from <200 https://sofifa.com/players?col=oa&sort=desc&offset=16500>
{'player_url': '/player/251978'}
2021-08-24 07:01:54 [scrapy.core.scrape] DEBUG: Scraped from <200 https://sofifa.com/players?col=oa&sort=desc&offset=16500>
{'player_url': '/player/254282'}
2021-08-24 07:01:54 [scrapy.core.scrape] DEBUG: Scraped from <200 https://sofifa.com/players?col=oa&sort=desc&offset=16500>
{'player_url': '/player/258890'}
2021-08-24 07:01:54 [scrapy.core.scrape] DEBUG: Scraped from <200 https://sofifa.com/players?col=oa&sort=desc&offset=16500>
{'player_url': '/player/260682'}
2021-08-24 07:01:54 [scrapy.core.scrape] DEBUG: Scraped from <200 https://sofifa.com/players?col=oa&sort=desc&offset=16500>
{'player_url': '/player/234571'}
2021-08-24 07:01:54 [scrapy.core.scrape] DEBUG: Scraped from <200 https://sofifa.com/players?col=oa&sort=desc&offset=16500>
{'player_url': '/player/257099'}
2021-08-24 07:01:54 [scrapy.core.scrape] DEBUG: Scraped from <200 https://sofifa.com/players?col=oa&sort=desc&offset=16500>
{'player_url': '/player/258635'}
2021-08-24 07:01:54 [scrapy.core.scrape] DEBUG: Scraped from <200 https://sofifa.com/players?col=oa&sort=desc&offset=16500>
{'player_url': '/player/246092'}
2021-08-24 07:01:54 [scrapy.core.scrape] DEBUG: Scraped from <200 https://sofifa.com/players?col=oa&sort=desc&offset=16500>
```

Bước 3: Dùng các ID đã thu thập từ file json để thu thập thông tin chi tiết của mỗi cầu thủ

sofifa.com/player/158023

REAL OVERALL RATING



Lionel Andrés Messi Cuccittini
RW ST CF 33y.o. (Jun 24, 1987) 170cm 72kg

Overall Rating: 93 | Potential: 93 | Value: €103.5M | Wage: €320K

PROFILE

Preferred Foot: Left | #Dribbler
Weak Foot: 4 ★ | #Distance Shooter
Skill Moves: 4 ★ | #FK Specialist
International Reputation: 5 ★ | #Acrobat
Work Rate: Medium/ Low | #Clinical Finisher
Body Type: Unique | #Complete Forward
Real Face: Yes
Release Clause: €191.5M
ID: 158023

PLAYER SPECIALITIES

RW | Position: RW | Jersey Number: 30 | Joined: Aug 10, 2021 | Contract Valid Until: 2023

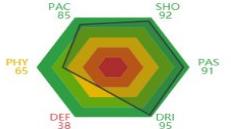
PARIS SAINT-GERMAIN | ARGENTINA

LIKES (1461) | DISLIKES (466) | FOLLOW (1392) | HISTORY VERSION (595)

LAYOUT 1 | 2 | 3

Best Position: RW | Best Overall Rating: 93

Best Position: RW | Best Overall Rating: 93

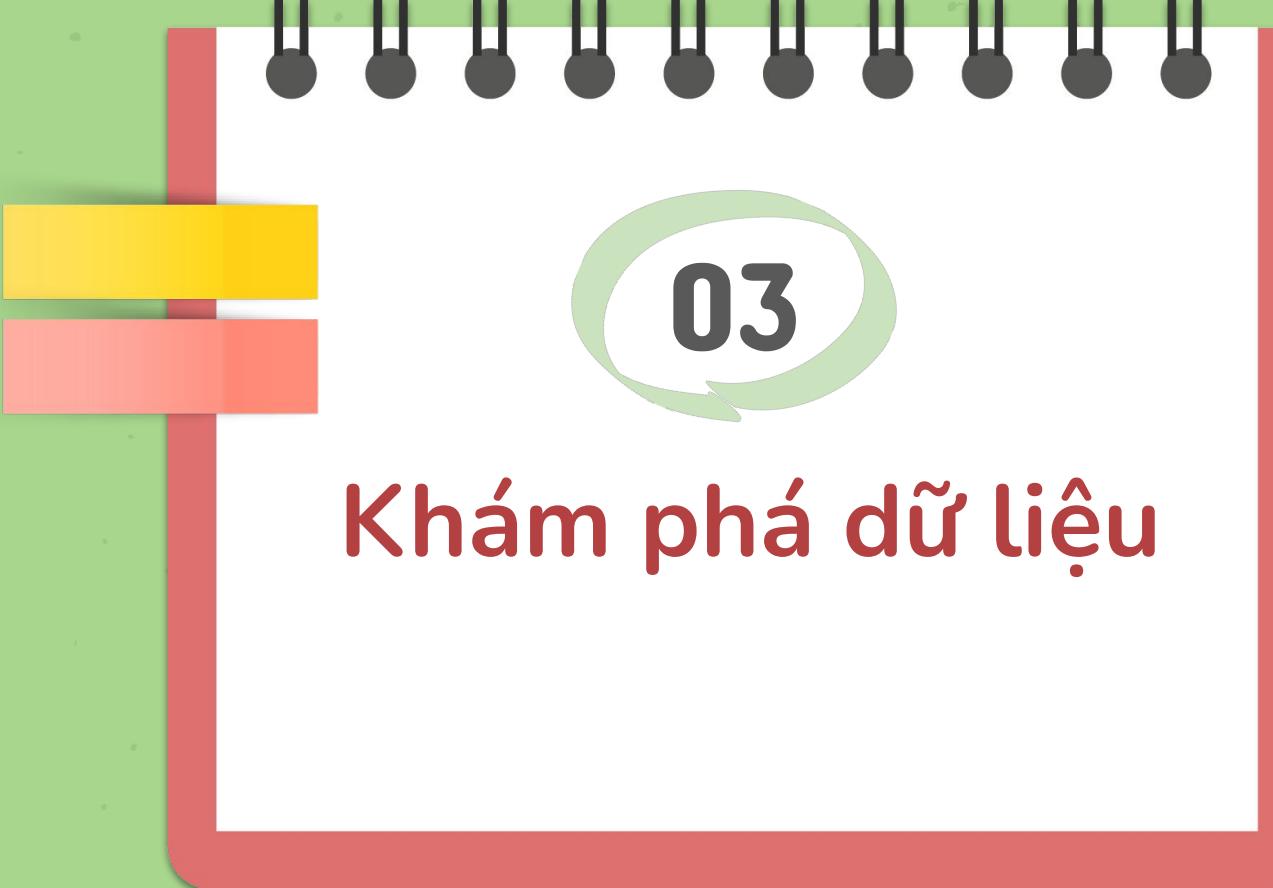


SIMILAR PLAYERS

ATTACKING	SKILL	MOVEMENT	POWER
85 Crossing 95 Finishing 70 Heading Accuracy 91 Short Passing 88 Volleys	96 Dribbling 93 Curve 94 FK Accuracy 91 Long Passing 96 Ball Control	91 Acceleration 80 Sprint Speed 91 Agility 94 Reactions 95 Balance	86 Shot Power 68 Jumping 72 Stamina 69 Strength 94 Long Shots
MENTALITY	DEFENDING	GOALKEEPING	TRAITS
44 Aggression 40 Interceptions 93 Positioning 95 Vision 75 Penalties 96 Composure	32 Defensive Awareness 35 Standing Tackle 24 Sliding Tackle	6 GK Diving 11 GK Handling 15 GK Kicking 14 GK Positioning 8 GK Reflexes	Finesse Shot Long Shot Taker (AI) Speed Dribbler (AI) Playmaker (AI) Outside Foot Shot One Club Player Team Player Chip Shot (AI)

Bước 4: Lưu thông tin chi tiết của mỗi cầu thủ đã thu thập được vào một file json khác

```
{'id': '261410', 'name': 'Leider Sebastián Berdugo Ruiz', 'short_name': 'L. Berdugo', 'photo_url': 'https://cdn.sofifa.com/players/261/410/21_120.png', 'primary_position': 'CAM', 'positions': ['CM'], 'age': '18', 'birth_date': '2002-08-23 12:17:27 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://sofifa.com/player/242211?units=mks> (referer: https://sofifa.com/player/261410?units=mks) player_count17931' 2021-08-23 12:17:27 [scrapy.core.scrapers] DEBUG: Scraped from <200 https://sofifa.com/player/242211?units=mks> {'id': '242211', 'name': 'Henry Woods', 'short_name': 'H. Woods', 'photo_url': 'https://cdn.sofifa.com/players/242/211/21_120.png', 'primary_position': 'RM', 'positions': ['CM', 'CAM', 'RB'], 'age': '20', 'birth_date': '1999/Sep/7' 2021-08-23 12:17:27 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://sofifa.com/player/259619?units=mks> (referer: https://sofifa.com/player/242211?units=mks) player_count17932' 2021-08-23 12:17:27 [scrapy.core.scrapers] DEBUG: Scraped from <200 https://sofifa.com/player/259619?units=mks> {'id': '259619', 'name': 'Jeyson Chura', 'short_name': 'J. Chura', 'photo_url': 'https://cdn.sofifa.com/players/259/619/21_120.png', 'primary_position': 'ST', 'positions': ['ST', 'RW'], 'age': '18', 'birth_date': '2002/Feb/3', 'height': '180cm' 2021-08-23 12:17:28 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://sofifa.com/player/262182?units=mks> (referer: https://sofifa.com/player/259619?units=mks) player_count17933' 2021-08-23 12:17:28 [scrapy.core.scrapers] DEBUG: Scraped from <200 https://sofifa.com/player/262182?units=mks> {'id': '262182', 'name': 'Sivert Øverby', 'short_name': 'S. Øverby', 'photo_url': 'https://cdn.sofifa.com/players/262/182/21_120.png', 'primary_position': 'CB', 'positions': ['CB'], 'age': '21', 'birth_date': '1999/Jun/10', 'height': '185cm' 2021-08-23 12:17:28 [scrapy.extensions.logstats] INFO: Crawled 17934 pages (at 103 pages/min), scraped 17933 items (at 103 items/min) 2021-08-23 12:17:28 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://sofifa.com/player/255014?units=mks> (referer: https://sofifa.com/player/262182?units=mks) player_count17934' 2021-08-23 12:17:28 [scrapy.core.scrapers] DEBUG: Scraped from <200 https://sofifa.com/player/255014?units=mks> {'id': '255014', 'name': 'Teddy Bartouche-Selbonne', 'short_name': 'T. Bartouche-Selbonne', 'photo_url': 'https://cdn.sofifa.com/players/255/014/21_120.png', 'primary_position': 'GK', 'positions': ['GK'], 'age': '23', 'birth_date': '2000/Mar/10' 2021-08-23 12:17:29 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://sofifa.com/player/256551?units=mks> (referer: https://sofifa.com/player/255014?units=mks) player_count17935' 2021-08-23 12:17:29 [scrapy.core.scrapers] DEBUG: Scraped from <200 https://sofifa.com/player/256551?units=mks> {'id': '256551', 'name': 'Joaquín Ignacio Gutiérrez Jara', 'short_name': 'J. Gutiérrez', 'photo_url': 'https://cdn.sofifa.com/players/256/551/21_120.png', 'primary_position': 'RB', 'positions': ['RB'], 'age': '17', 'birth_date': '2005/Mar/10' 2021-08-23 12:17:29 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://sofifa.com/player/261159?units=mks> (referer: https://sofifa.com/player/256551?units=mks) player_count17936' 2021-08-23 12:17:30 [scrapy.core.scrapers] DEBUG: Scraped from <200 https://sofifa.com/player/261159?units=mks> {'id': '261159', 'name': 'Martijn Beernaert', 'short_name': 'M. Beernaert', 'photo_url': 'https://cdn.sofifa.com/players/261/159/21_120.png', 'primary_position': 'GK', 'positions': ['GK'], 'age': '17', 'birth_date': '2002/Oct/3', 'height': '188cm' 2021-08-23 12:17:30 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://sofifa.com/player/257832?units=mks> (referer: https://sofifa.com/player/261159?units=mks) player_count17937' 2021-08-23 12:17:30 [scrapy.core.scrapers] DEBUG: Scraped from <200 https://sofifa.com/player/257832?units=mks> {'id': '257832', 'name': 'Bradley Foster-Thenigen', 'short_name': 'B. Foster', 'photo_url': 'https://cdn.sofifa.com/players/257/832/21_120.png', 'primary_position': 'GK', 'positions': ['GK'], 'age': '18', 'birth_date': '2001/Oct/5' 2021-08-23 12:17:31 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://sofifa.com/player/256553?units=mks> (referer: https://sofifa.com/player/257832?units=mks) player_count17938' 2021-08-23 12:17:31 [scrapy.core.scrapers] DEBUG: Scraped from <200 https://sofifa.com/player/256553?units=mks> {'id': '256553', 'name': 'Álvaro Sebastián Garrido Podlech', 'short_name': 'A. Garrido', 'photo_url': 'https://cdn.sofifa.com/players/256/553/21_120.png', 'primary_position': 'CAM', 'positions': ['CM'], 'age': '20', 'birth_date': '2001/Mar/10' 2021-08-23 12:17:31 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://sofifa.com/player/257065?units=mks> (referer: https://sofifa.com/player/256553?units=mks) player_count17939' 2021-08-23 12:17:31 [scrapy.core.scrapers] DEBUG: Scraped from <200 https://sofifa.com/player/257065?units=mks> {'id': '257065', 'name': 'Nico Lemoine', 'short_name': 'N. Lemoine', 'photo_url': 'https://cdn.sofifa.com/players/257/065/21_120.png', 'primary_position': 'RM', 'positions': ['RM', 'RW'], 'age': '20', 'birth_date': '2000/Apr/10', 'height': '185cm' 2021-08-23 12:17:31 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://sofifa.com/player/257066?units=mks> (referer: https://sofifa.com/player/257065?units=mks) player_count17940'
```



03

Khám phá dữ liệu

Dữ liệu có 18922 dòng và 62 cột. Dưới đây là thông tin của một số cột

#	Column	Non-Null Count	Dtype
0	id	18922	non-null
1	name	18922	non-null
2	short_name	18922	non-null
3	photo_url	18922	non-null
4	primary_position	18922	non-null
5	positions	18922	non-null
6	age	18922	non-null
7	birth_date	18922	non-null
8	height	18922	non-null
9	weight	18922	non-null
10	Overall Rating	18922	non-null
11	Potential	18922	non-null
12	Value	18922	non-null
13	Wage	18922	non-null
14	Preferred Foot	18922	non-null
15	Weak Foot	18922	non-null
16	Skill Moves	18922	non-null
17	International Reputation	18922	non-null
18	Work Rate	18922	non-null
19	Body Type	18922	non-null
20	Real Face	18922	non-null
21	Release Clause	17322	non-null
22	teams	18922	non-null
23	player_traits	18922	non-null
24	player_specialities	18922	non-null
25	nationality	18922	non-null
26	club	18922	non-null
27	nation_club	18922	non-null
28	Crossing	18922	non-null
29	Finishing	18922	non-null
30	HeadingAccuracy	18922	non-null
31	ShortPassing	18922	non-null
32	Volleyes	18922	non-null
33	Dribbling	18922	non-null
34	Curve	18922	non-null
35	FK_Accuracy	18922	non-null
36	Long_Passing	18922	non-null
37	Ball_Control	18922	non-null
38	Acceleration	18922	non-null
39	Sprint_Speed	18922	non-null
40	Agility	18922	non-null
41	Reactions	18922	non-null
42	Balance	18922	non-null
43	Shot_Power	18922	non-null
44	Jumping	18922	non-null
45	Stamina	18922	non-null

Dữ liệu chỉ có duy nhất một cột 'Release Clause' là
bị thiếu 1600 giá trị

Real Face	0
Release Clause	1600
teams	0
player_traits	0
player_specialities	0
nationality	0
club	0
nation_club	0
Crossing	0
Finishing	0
HeadingAccuracy	0
ShortPassing	0
Volleys	0
Dribbling	0
Curve	0
FK Accuracy	0

Vì hai cột 'Value' và 'Wage' có chứa các ký tự viết tắt nên chúng ta sẽ chuyển về các giá trị đầy đủ

Value	Wage
€58.5M	€135K
€39.5M	€80K
€28M	€70K
€30M	€71K
€24M	€115K
...	...
€190K	€1K
€170K	€3K
€210K	€2K
€190K	€2K
€140K	€4K

Value	Wage
103500000.0	560000.0
124500000.0	270000.0
63000000.0	220000.0
120000000.0	125000.0
127500000.0	370000.0

Dưới đây là tổng số lượng cầu thủ theo từng vị trí và
theo câu lạc bộ

	Positions	Total Players
0	CB	3689
1	ST	2678
2	CAM	2272
3	GK	2098
4	RM	1491
5	CDM	1459
6	CM	1084
7	RB	1030
8	LB	1015
9	LM	897
10	RWB	356
11	RW	300
12	LWB	295
13	LW	192
14	CF	66

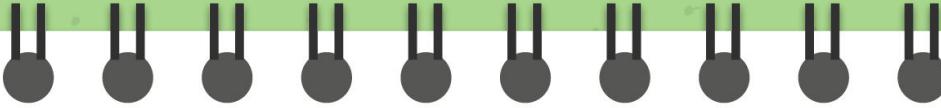
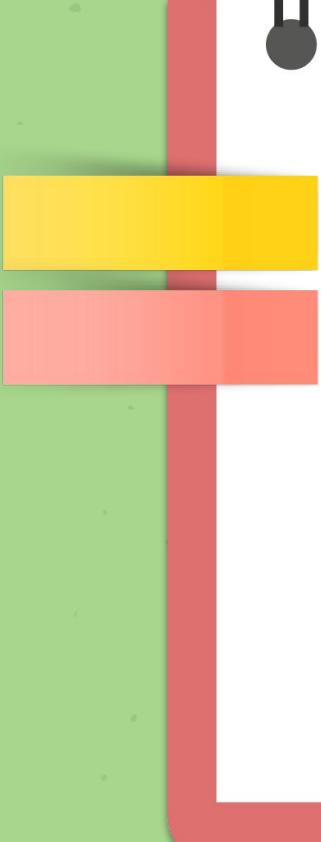
	Club	Total Players
0	West Ham United	33
1	Parma	33
2	FC Barcelona	33
3	Angers SCO	33
4	Roma	33
...
678	Shelbourne FC	19
679	Club Athletico Paranaense	19
680	Corinthians	19
681	Jiangsu Suning FC	18
682	Palmeiras	18

Dưới đây là tổng số lượng cầu thủ theo quốc tịch và theo đội tuyển quốc gia

	Nationality	Total Players
0	England	1722
1	Germany	1189
2	Spain	1075
3	France	1025
4	Argentina	915
...
158	Saint Lucia	1
159	Vietnam	1
160	Guam	1
161	Aruba	1
162	Papua New Guinea	1

	Nation Club	Total Players
0	France	23
1	Wales	23
2	Germany	23
3	Canada	23
4	Finland	23
5	Chile	23
6	Portugal	23
7	Belgium	23
8	Peru	23
9	Sweden	23
10	Poland	23
11	Republic of Ireland	23
12	Iceland	23
13	Mexico	23
14	Turkey	23
15	Netherlands	23
16	Ecuador	23
17	Northern Ireland	23
18	Cameroon	23
19	United States	23
20	Argentina	23
21	Australia	23
22	Austria	23
23	Denmark	23
24	Spain	23
25	Venezuela	23
26	New Zealand	23
27	Brazil	23
28	Colombia	23
29	India	23
30	Uruguay	23
31	Paraguay	23

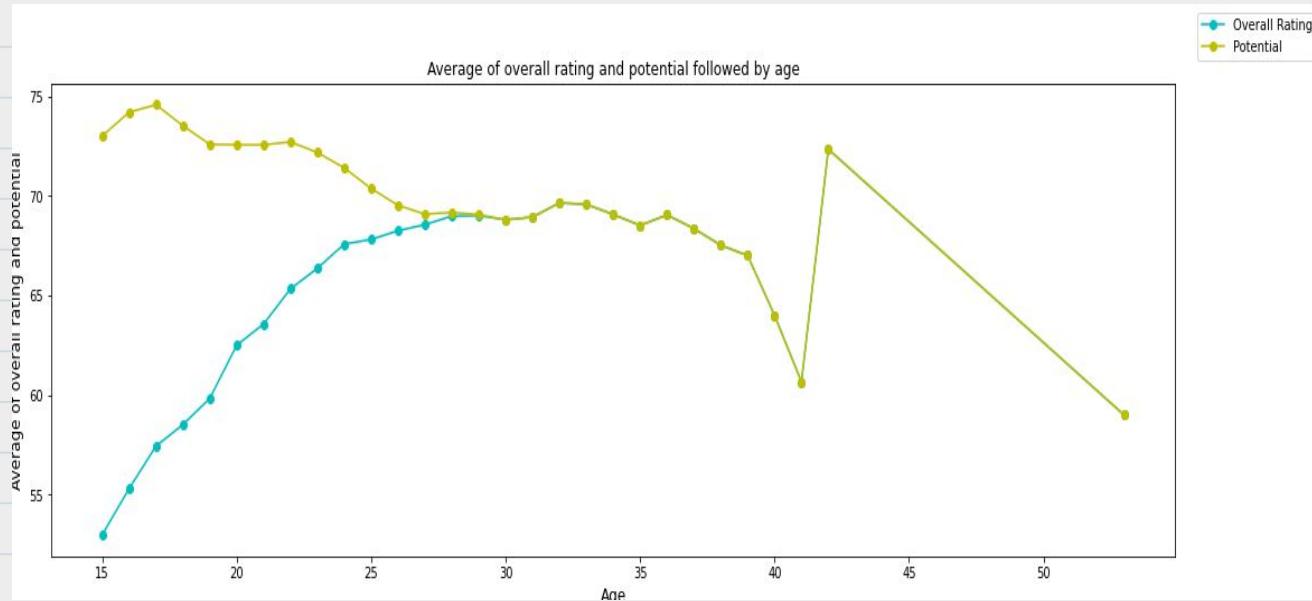
- Giá trị trung bình cao nhất là K.Mbappé: €185.500.000
- Lương trung bình cao nhất là L.Messi: €560.000
- Chỉ số tổng quát cao nhất là L.Messi: 93
- Chỉ số tiềm năng cao nhất là K.Mbappé: 95
- Cầu thủ có cân nặng lớn nhất là A.Akinfenwa: 110kg
- Cầu thủ có chiều cao lớn nhất là T.Holý: 206cm
- Cầu thủ có phí giải phóng cao nhất là K.Mappé: €357.000.000
- Cầu thủ lớn tuổi nhất là K.Miura: 53



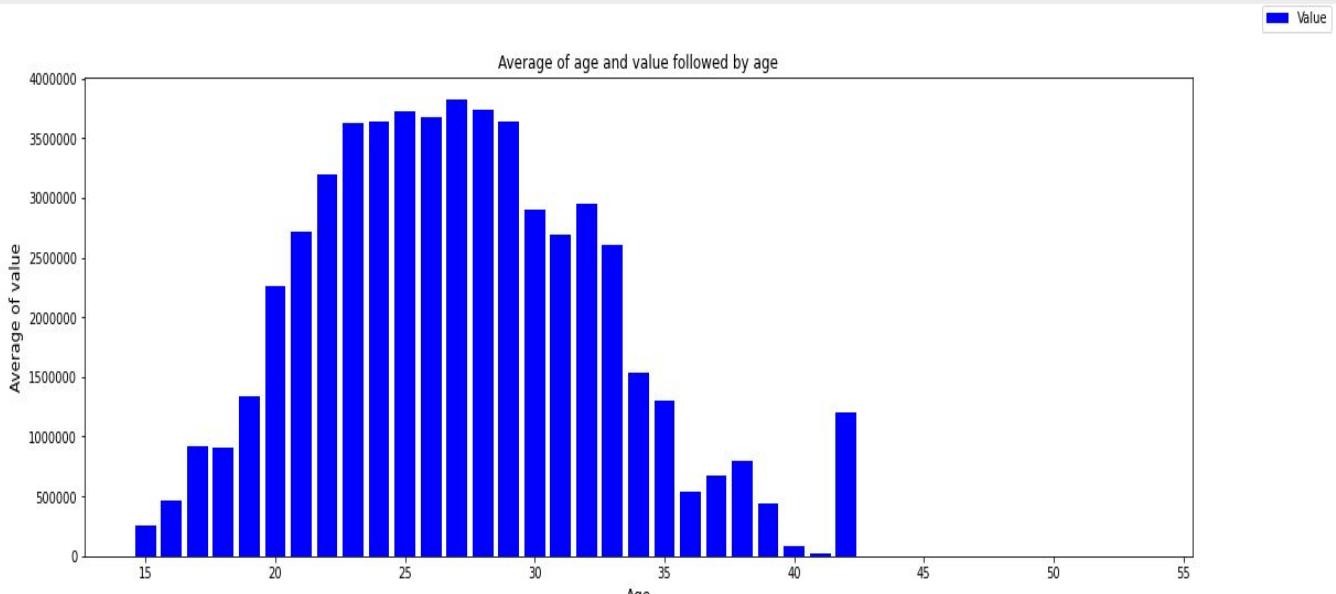
04

Trực quan hóa dữ liệu

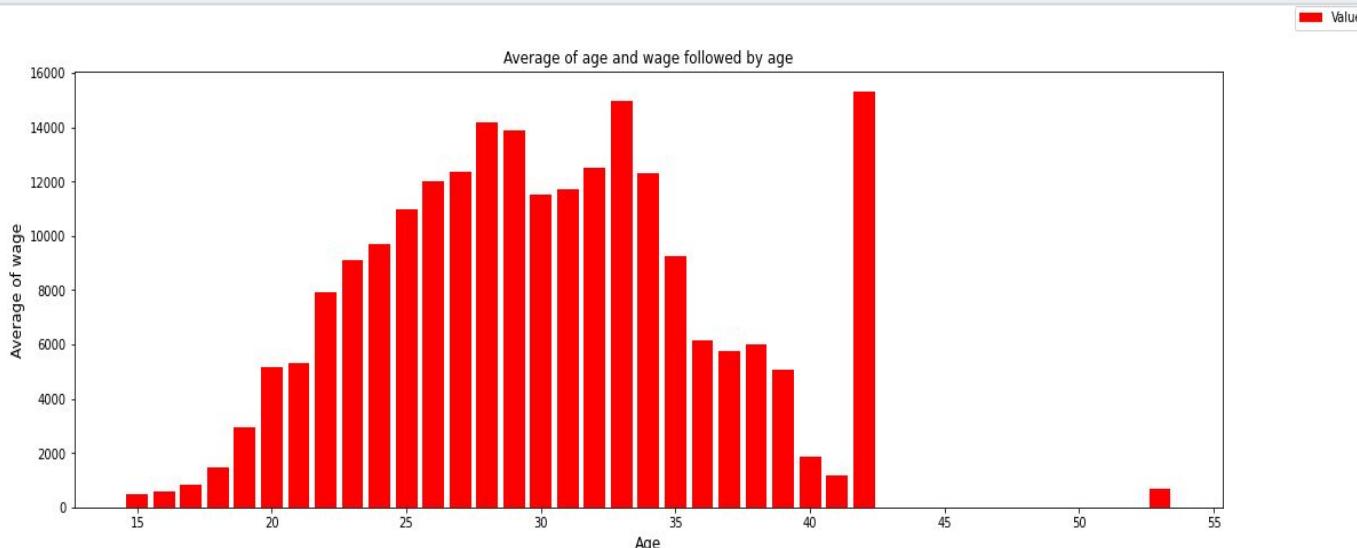
Với các cầu thủ dưới 30 tuổi thì trung bình chỉ số tổng quát và tiềm năng khá cao nhưng khi trên 30 tuổi thì có phần chững lại do phong độ đã đạt đỉnh cao nhất



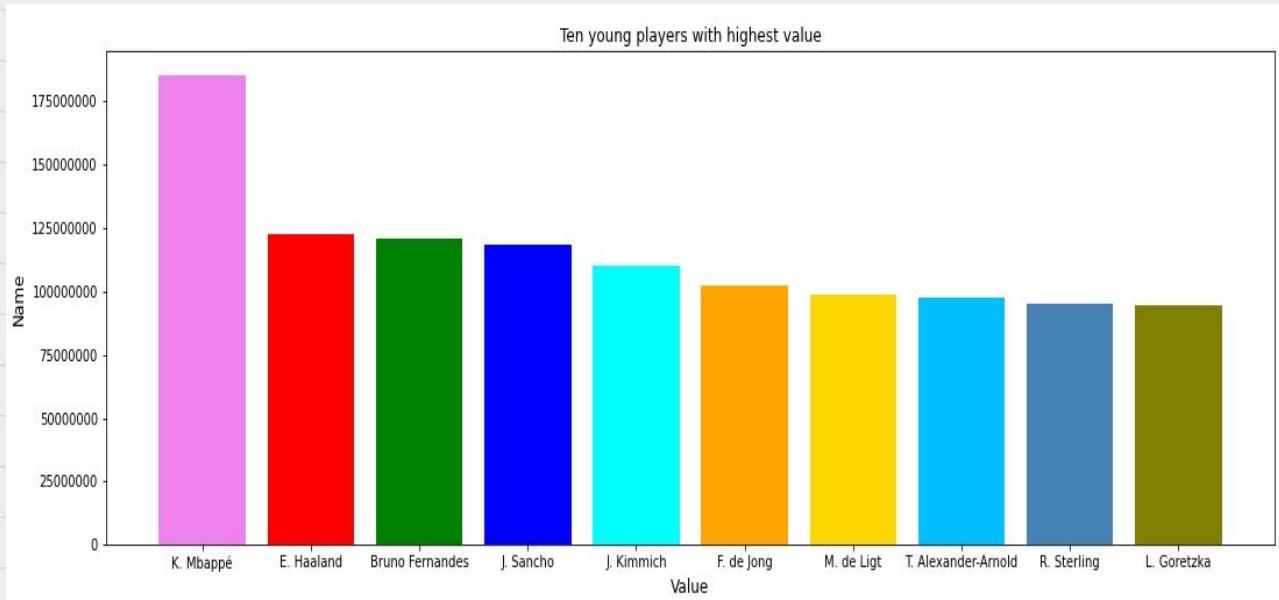
Một cách tương tự, với các cầu thủ dưới 30 tuổi thì trung bình giá trị tăng khá cao nhưng khi trên 30 tuổi thì giá trị bị giảm một cách nhanh chóng



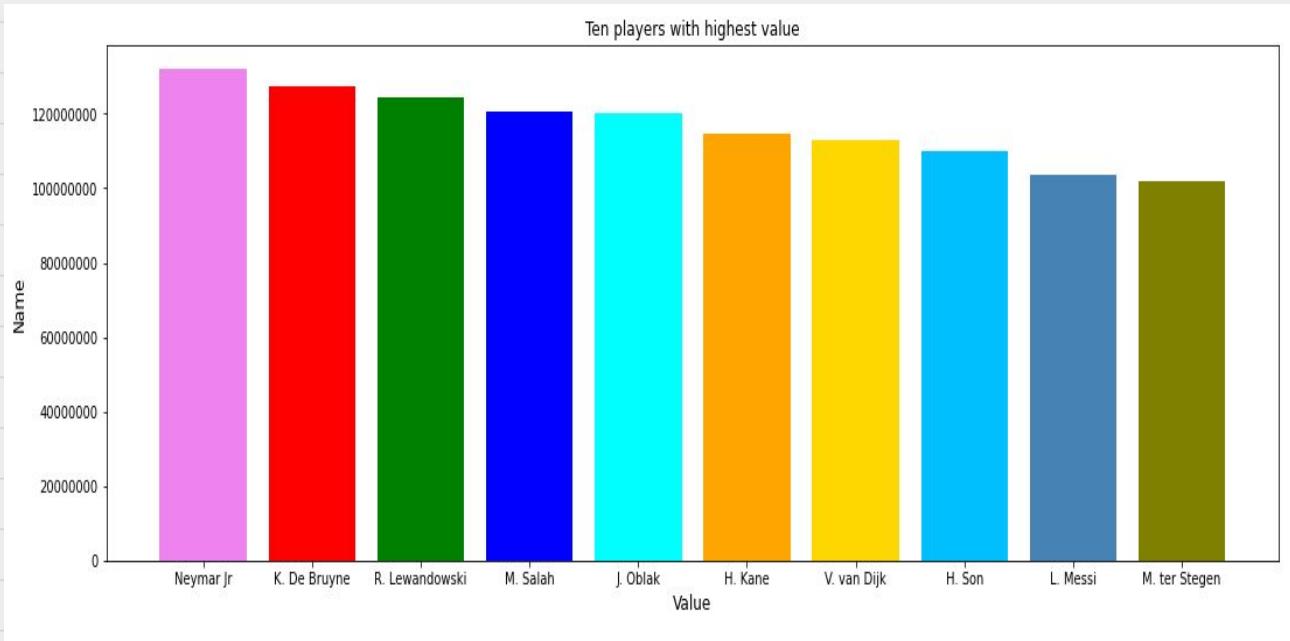
Về lương thì các cầu thủ dưới 30 tuổi mức lương tăng
khá nhanh và càng giảm khi trên 30 tuổi nhưng cũng có
một vài cầu thủ do vẫn giữ được phong độ thi đấu cao
nên mức lương khi ngoài 30 tuổi vẫn rất cao



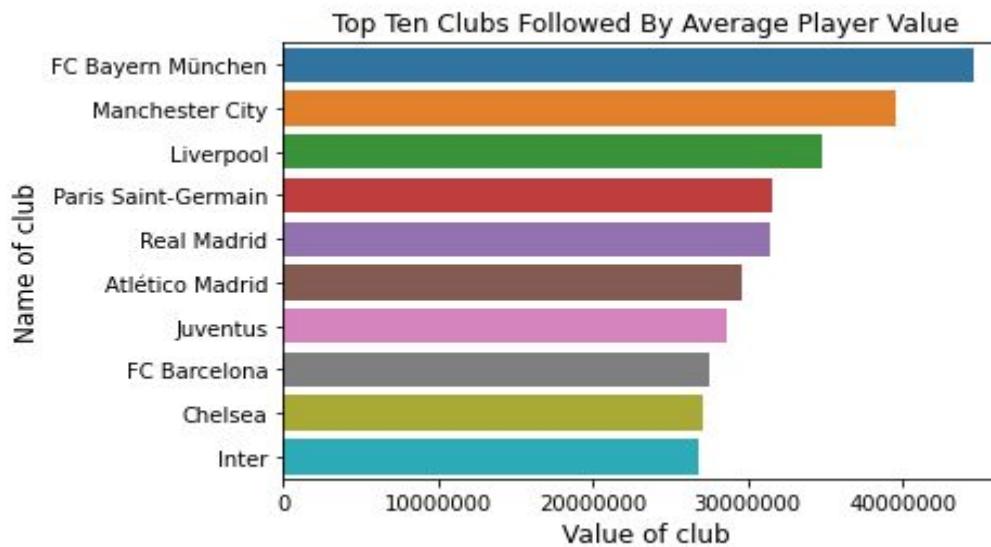
Dưới đây là top 10 các cầu thủ trẻ (từ 25 tuổi trở xuống) mà có giá trị cao nhất



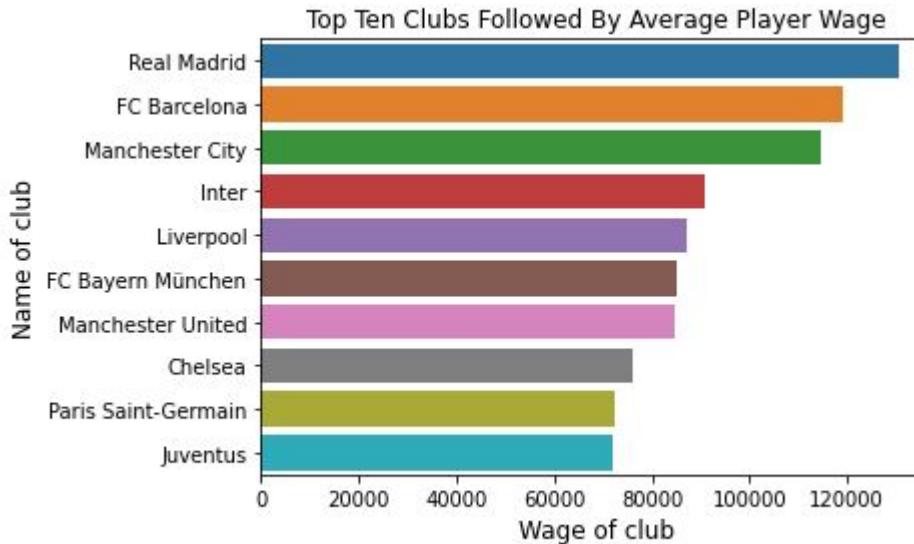
Dưới đây là top 10 các cầu thủ (trên 25 tuổi) mà có giá trị cao nhất



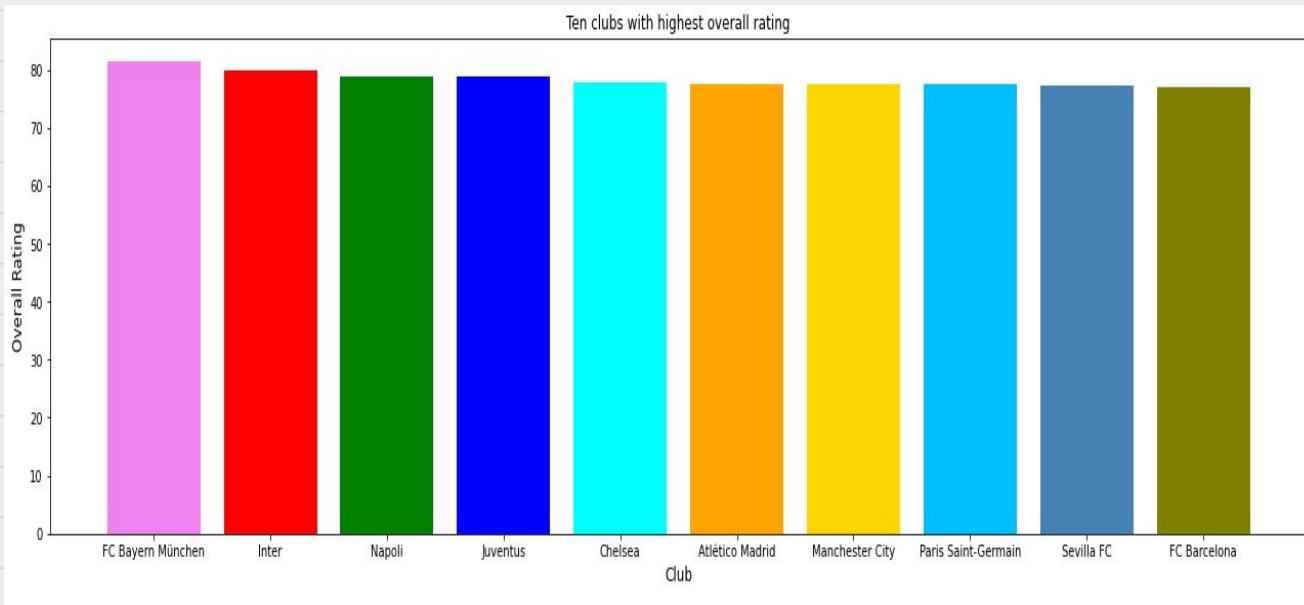
Dưới đây là top 10 các câu lạc bộ xếp theo trung bình giá trị của các cầu thủ thuộc câu lạc bộ đó



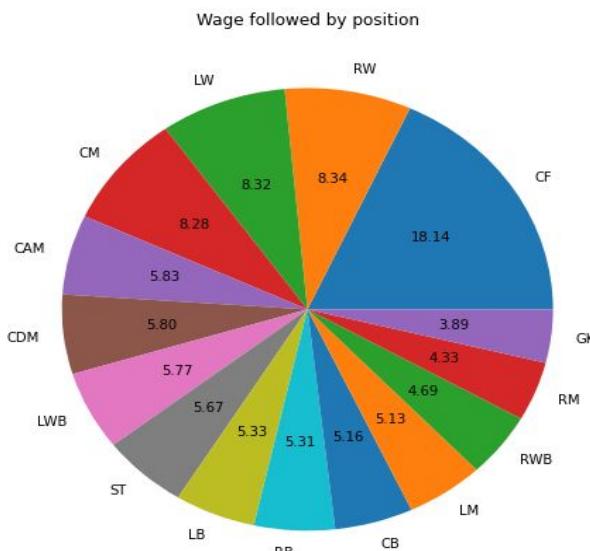
Dưới đây là top 10 các câu lạc bộ xếp theo trung bình
lương của các cầu thủ thuộc câu lạc bộ đó



Dưới đây là top 10 các câu lạc bộ xếp theo trung bình chỉ số tổng quát của các cầu thủ thuộc câu lạc bộ đó



Dưới đây là trung bình mức lương của các cầu thủ theo từng vị trí cụ thể



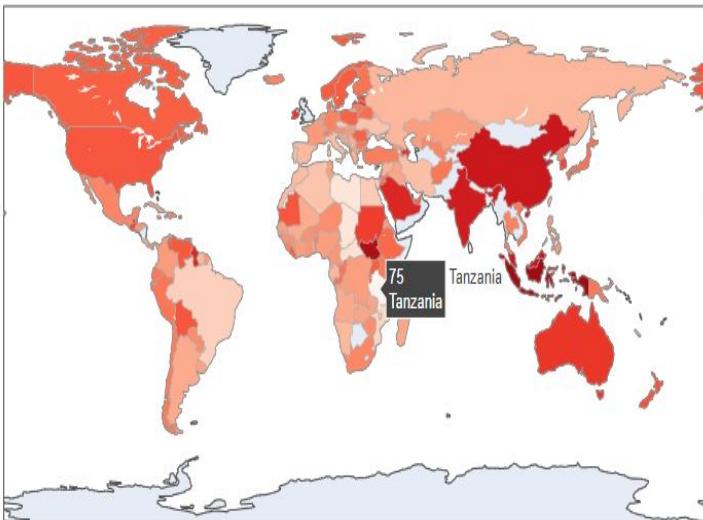
Dưới đây là danh sách các cầu thủ có chỉ số tổng quát cao nhất theo từng vị trí cụ thể

	short_name	primary_position	age	Overall	Rating	Potential	Value	Wage	club	nationality
0	L. Messi	RW	33	93	93	103500000.0	560000.0	FC Barcelona	Argentina	
1	R. Lewandowski	ST	31	92	92	124500000.0	270000.0	FC Bayern München	Poland	
2	Cristiano Ronaldo	ST	35	92	92	63000000.0	220000.0	Juventus	Portugal	
3	J. Oblak	GK	27	91	93	120000000.0	125000.0	Atlético Madrid	Slovenia	
4	K. De Bruyne	CM	29	91	91	127500000.0	370000.0	Manchester City	Belgium	
5	Neymar Jr	LW	28	91	91	132000000.0	270000.0	Paris Saint-Germain	Brazil	
8	V. van Dijk	CB	28	90	91	113000000.0	210000.0	Liverpool	Netherlands	
12	J. Kimmich	CDM	25	89	90	110000000.0	150000.0	FC Bayern München	Germany	
14	Casemiro	CDM	28	89	89	90500000.0	310000.0	Real Madrid	Brazil	
15	H. Son	LM	27	89	89	110000000.0	210000.0	Tottenham Hotspur	Korea Republic	
18	K. Benzema	CF	32	89	89	83500000.0	350000.0	Real Madrid	France	
21	Bruno Fernandes	CAM	25	88	91	121000000.0	240000.0	Manchester United	Portugal	
28	A. Robertson	LB	26	87	88	85500000.0	155000.0	Liverpool	Scotland	
47	T. Alexander-Arnold	RB	21	86	91	97500000.0	100000.0	Liverpool	England	
53	S. Gnabry	RM	24	86	88	88000000.0	110000.0	FC Bayern München	Germany	
54	Carvajal	RB	28	86	86	61500000.0	230000.0	Real Madrid	Spain	
133	A. Wan-Bissaka	RWB	22	83	87	49500000.0	115000.0	Manchester United	England	
134	F. Mendi	LWB	25	83	88	49500000.0	160000.0	Real Madrid	France	
141	Angelino	LWB	23	83	86	46000000.0	58000.0	RB Leipzig	Spain	
150	Gayà	LWB	25	83	88	49500000.0	46000.0	Valencia CF	Spain	

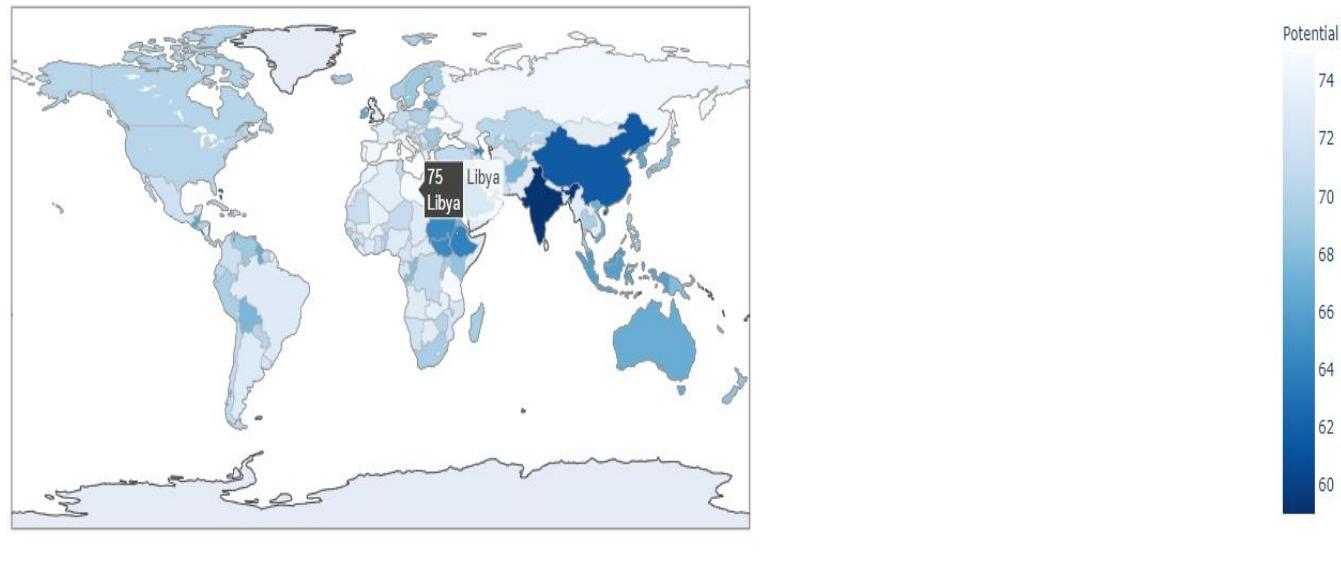
Dưới đây là các chỉ số xếp theo mức độ quan trọng nhất của các cầu thủ theo từng vị trí cụ thể

primary_position	Attacking	Skill	Movement	Power	Mentality	Defending	Goalkeeping
CAM	57.198415	63.238820	69.891285	60.096303	56.864510	43.578638	10.275704
CB	46.252372	45.643806	57.860287	59.264787	52.054622	65.798139	10.459854
CDM	54.254421	58.751337	63.688965	64.925840	60.912497	64.911355	10.510761
CF	66.436364	67.087879	72.472727	68.336364	61.851010	33.777778	10.163636
CM	58.494280	64.055904	66.179151	65.487454	62.919434	61.094096	10.492804
GK	15.494280	17.427836	42.680553	41.019828	25.092628	13.795996	63.879123
LB	52.395862	55.937537	68.785616	60.169458	56.206732	63.523810	10.548966
LM	57.606912	61.153400	72.154961	61.069342	55.006132	41.238201	10.338907
LW	60.314583	62.586458	74.852083	62.050000	55.866319	34.043403	10.439583
LWB	52.649492	57.594576	70.452881	61.088814	56.601130	62.783051	10.413559
RB	51.857087	54.167961	69.098447	59.705243	55.794337	63.723301	10.482524
RM	56.030852	59.120858	72.018109	60.072032	53.851218	39.825844	10.375050
RW	60.355333	61.822000	75.688667	62.921333	56.111667	34.388889	10.431333
RWB	51.943258	55.496629	70.639888	60.656180	56.180712	62.501873	10.267978
ST	59.525990	54.028977	65.927110	66.094100	54.335574	26.293129	10.453174

Dưới đây là bản đồ thể hiện trung bình chỉ số tổng quát của các cầu thủ theo quốc gia thì Tanzania là nước có trung bình chỉ số tổng quát cao nhất



Dưới đây là bản đồ thể hiện trung bình tiềm năng của các cầu thủ theo quốc gia thì Libya là nước có trung bình tiềm năng cao nhất





05

Tiền xử lý dữ liệu

Đầu tiên, chúng ta sẽ tách các thuộc tính cầu thủ chứa trong các dict như hình bên dưới thành các cột riêng lẻ và với cột 'Release Clause' có 1600 giá trị thiếu chúng ta sẽ thay thế toàn bộ là 0

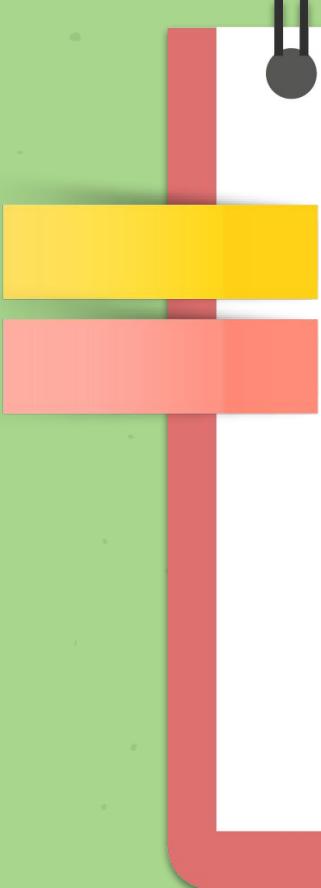
attacking	skill	movement	power	mentality	defending	goalkeeping
{'Crossing': 85, 'Finishing': 95, 'HeadingAccuracy': 93, 'FKAccuracy': 9...}	{'Dribbling': 96, 'Curve': 93, 'FKAccuracy': 9...}	{'Acceleration': 91, 'SprintSpeed': 80, 'Agility': 72...}	{'ShotPower': 86, 'Jumping': 68, 'Stamina': 72...}	{'Aggression': 44, 'Interceptions': 40, 'Positioning': 38...}	{'DefensiveAwareness': 32, 'StandingTackle': 30, 'SlidingTackle': 28...}	{'GKDivi...': 6, 'GKHandling': 11, 'GKKicking': 10...}
{'Crossing': 71, 'Finishing': 95, 'HeadingAccuracy': 79, 'FKAccuracy': 8...}	{'Dribbling': 85, 'Curve': 79, 'FKAccuracy': 8...}	{'Acceleration': 77, 'SprintSpeed': 79, 'Agility': 76...}	{'ShotPower': 90, 'Jumping': 84, 'Stamina': 76...}	{'Aggression': 81, 'Interceptions': 49, 'Positioning': 47...}	{'DefensiveAwareness': 35, 'StandingTackle': 40, 'SlidingTackle': 35...}	{'GKDivi...': 15, 'GKHandling': 6, 'GKKicking': 10...}
{'Crossing': 84, 'Finishing': 95, 'HeadingAccuracy': 81, 'FKAccuracy': 7...}	{'Dribbling': 88, 'Curve': 81, 'FKAccuracy': 7...}	{'Acceleration': 87, 'SprintSpeed': 91, 'Agility': 84...}	{'ShotPower': 94, 'Jumping': 95, 'Stamina': 84...}	{'Aggression': 63, 'Interceptions': 29, 'Positioning': 27...}	{'DefensiveAwareness': 28, 'StandingTackle': 30, 'SlidingTackle': 28...}	{'GKDivi...': 7, 'GKHandling': 11, 'GKKicking': 10...}
{'Crossing': 13, 'Finishing': 11, 'HeadingAccuracy': 13, 'FKAccuracy': 1...}	{'Dribbling': 12, 'Curve': 13, 'FKAccuracy': 1...}	{'Acceleration': 43, 'SprintSpeed': 60, 'Agility': 41...}	{'ShotPower': 59, 'Jumping': 78, 'Stamina': 41...}	{'Aggression': 34, 'Interceptions': 19, 'Positioning': 18...}	{'DefensiveAwareness': 27, 'StandingTackle': 1, 'SlidingTackle': 1...}	{'GKDivi...': 87, 'GKHandling': 92, 'GKKicking': 10...}
{'Crossing': 94, 'Finishing': 82, 'HeadingAccuracy': 85, 'FKAccuracy': 8...}	{'Dribbling': 88, 'Curve': 77, 'FKAccuracy': 8...}	{'Acceleration': 77, 'SprintSpeed': 77, 'Agility': 89...}	{'ShotPower': 91, 'Jumping': 63, 'Stamina': 89...}	{'Aggression': 76, 'Interceptions': 66, 'Positioning': 64...}	{'DefensiveAwareness': 68, 'StandingTackle': 6, 'SlidingTackle': 6...}	{'GKDivi...': 15, 'GKHandling': 13, 'GKKicking': 10...}

Dưới đây là các cột thể hiện các thuộc tính cụ thể
của từng cầu thủ sau khi chúng ta đã tách từ các dict

Crossing	Finishing	HeadingAccuracy	ShortPassing	Volleys	Dribbling	Curve	FK_Accuracy	Long_Passing	Ball_Control	Acceleration	Sprint_Speed	Agility	Reactions	Balance	Shot_Power	Jumping	Stamina	Strength	Long_Shots	Aggres
85	95	70	91	88	96	93	94	91	96	91	80	91	94	95	86	68	72	69	94	
71	95	90	85	89	85	79	85	70	88	77	79	77	93	82	90	84	76	86	87	
84	95	90	82	86	88	81	76	77	92	87	91	87	95	71	94	95	84	78	93	
13	11	15	43	13	12	13	14	40	30	43	60	67	88	49	59	78	41	78	12	
94	82	55	94	82	88	85	83	93	92	77	77	79	91	78	91	63	89	74	91	

Thứ hai, chúng ta sẽ tách các dict từ các cột quốc tịch, câu lạc bộ và đội tuyển quốc gia của các cầu thủ

nationality	club	nation_club	club	club_url	nationality	nationality_url	nation_club	nation_club_url
{'name': 'Argentina', 'url': 'https://cdn.sofifa.com/flags/ar.png'}	{'name': 'FC Barcelona', 'url': 'https://cdn.sofifa.com/teams/241/60.png'}	{'name': 'Argentina', 'url': 'https://cdn.sofifa.com/nations/1369/60.png'}	FC Barcelona	https://cdn.sofifa.com/teams/241/60.png	Argentina	https://cdn.sofifa.com/flags/ar.png	Argentina	https://cdn.sofifa.com/nations/1369/60.png
{'name': 'Poland', 'url': 'https://cdn.sofifa.com/flags/pl.png'}	{'name': 'FC Bayern München', 'url': 'https://cdn.sofifa.com/teams/21/60.png'}	{'name': 'Poland', 'url': 'https://cdn.sofifa.com/nations/1353/60.png'}	FC Bayern München	https://cdn.sofifa.com/teams/21/60.png	Poland	https://cdn.sofifa.com/flags/pl.png	Poland	https://cdn.sofifa.com/nations/1353/60.png
{'name': 'Portugal', 'url': 'https://cdn.sofifa.com/flags/pt.png'}	{'name': 'Juventus', 'url': 'https://cdn.sofifa.com/teams/45/60.png'}	{'name': 'Portugal', 'url': 'https://cdn.sofifa.com/nations/1354/60.png'}	Juventus	https://cdn.sofifa.com/teams/45/60.png	Portugal	https://cdn.sofifa.com/flags/pt.png	Portugal	https://cdn.sofifa.com/nations/1354/60.png
{'name': 'Slovenia', 'url': 'https://cdn.sofifa.com/flags/si.png'}	{'name': 'Atlético Madrid', 'url': 'https://cdn.sofifa.com/teams/240/60.png'}	{'name': 'Slovenia', 'url': 'https://cdn.sofifa.com/nations/1361/60.png'}	Atlético Madrid	https://cdn.sofifa.com/teams/240/60.png	Slovenia	https://cdn.sofifa.com/flags/si.png	Slovenia	https://cdn.sofifa.com/nations/1361/60.png
{'name': 'Belgium', 'url': 'https://cdn.sofifa.com/flags/be.png'}	{'name': 'Manchester City', 'url': 'https://cdn.sofifa.com/teams/10/60.png'}	{'name': 'Belgium', 'url': 'https://cdn.sofifa.com/nations/1325/60.png'}	Manchester City	https://cdn.sofifa.com/teams/10/60.png	Belgium	https://cdn.sofifa.com/flags/be.png	Belgium	https://cdn.sofifa.com/nations/1325/60.png



06

Mô hình hóa dữ liệu

Ở đây chúng ta sẽ đặt ra hai câu hỏi là với các chỉ số thuộc tính hiện tại của một cầu thủ như hình bên dưới thì chúng ta có thể dự đoán được cầu thủ đó có thể thi đấu được ở vị trí nào trên sân và chỉ số tổng quát của cầu thủ đó có thể đạt được là bao nhiêu

ATTACKING	SKILL	MOVEMENT	POWER
85 Crossing	96 Dribbling	91 Acceleration	86 Shot Power
95 Finishing	93 Curve	80 Sprint Speed	68 Jumping
70 Heading Accuracy	94 FK Accuracy	91 Agility	72 Stamina
91 Short Passing	91 Long Passing	94 Reactions	69 Strength
88 Volleys	96 Ball Control	95 Balance	94 Long Shots
MENTALITY	DEFENDING	GOALKEEPING	
44 Aggression	32 Defensive Awareness	6 GK Diving	
40 Interceptions	35 Standing Tackle	11 GK Handling	
93 Positioning	24 Sliding Tackle	15 GK Kicking	
95 Vision		14 GK Positioning	
75 Penalties		8 GK Reflexes	
96 Composure			

	LS	ST	RS	
	89+3	89+3	89+3	
LW	LF	CF	RF	RW
92+0	93+0	93+0	93+0	92+0
LAM	CAM	RAM		
93+0	93+0	93+0		
LM	LCM	CM	RCM	RM
91+2	87+3	87+3	87+3	91+2
LWB	LDM	CDM	RDM	RWB
66+3	65+3	65+3	65+3	66+3
LB	LCB	CB	RCB	RB
62+3	52+3	52+3	52+3	62+3
	GK			
	19+3			

Ở câu hỏi đầu tiên, chúng ta sẽ chia thành hai cách là dự đoán từng vị trí cụ thể cho các cầu thủ và gộp các vị trí thành bốn nhóm chính bao gồm: nhóm màu xanh cho các cầu thủ ở vị trí tấn công, nhóm màu đỏ cho các cầu thủ ở vị trí trung tâm, nhóm màu vàng cho các cầu thủ ở vị trí phòng ngự và nhóm còn lại cho các cầu thủ ở vị trí thủ môn

Ở cách môt, mô hình mà chúng ta sử dụng là logistic regression với max_iter = 1000, multi_class = 'multinomial', solver = 'lbfgs' và tỷ lệ train/test là 8/2. Độ chính xác khi chạy trên tập train là: 92% và trên tập test là: 91%. Dưới đây là hình minh họa vị trí ban đầu của cầu thủ và vị trí sau khi sử dụng mô hình để dự đoán

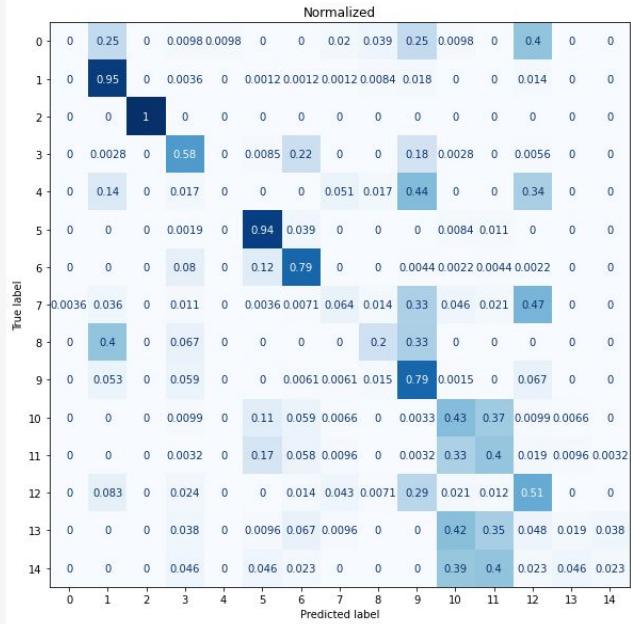
	primary_position						photo_url	primary_position
0	FW		0	Lionel Andrés Messi Cuccittini	L. Messi	https://cdn.sofifa.com/players/158/023/21_120.png	MF	
1	FW		1	Robert Lewandowski	R. Lewandowski	https://cdn.sofifa.com/players/188/545/21_120.png	FW	
2	FW		2	Cristiano Ronaldo dos Santos Aveiro	Cristiano Ronaldo	https://cdn.sofifa.com/players/020/801/21_120.png	FW	
3	GK		3	Jan Oblak	J. Oblak	https://cdn.sofifa.com/players/200/389/21_120.png	GK	
4	MF		4	Kevin De Bruyne	K. De Bruyne	https://cdn.sofifa.com/players/192/985/21_120.png	MF	

Ở cách hai, chúng ta sẽ chuyển 15 vị trí của toàn bộ cầu thủ trong dữ liệu thành các con số từ 0-14, mô hình mà nhóm sử dụng là RandomForestClassifier với random_state = 42, tỷ lệ train/test là: 7/3. Độ chính xác khi chạy trên tập train là: 100%, khi chạy trên tập test là: 71%. Dưới đây là hình minh họa kết quả dự đoán của mô hình với tổng số hàng dự đoán đúng là: 4021/5677

	Order	Name	Position	Predicted Position
0	8336	L. Schmitz	LB	CDM
1	5282	Botía	CB	CB
2	4525	M. Dupé	GK	GK
3	17374	I. Dahlqvist	RM	RM
4	9243	A. Mabaso	RB	LB

Tuy nhiên, có một số cầu thủ có thể chơi đa dạng các vị trí tại các tuyến tấn công, trung tâm hoặc phòng thủ nên khi so sánh kết quả các vị trí đã dự đoán với tất cả các vị trí mà một cầu thủ có thể chơi được thì kết quả tổng số hàng dự đoán đúng đã tăng từ 4021 lên 4198/5677. Dưới đây là hình minh họa kết quả dự đoán của mô hình so với toàn bộ các vị trí mà các cầu thủ có thể chơi

	Order	Name	Position	Predicted Position	All Positions Player Can Play
1	5282	Botía	CB	CB	['CB']
2	4525	M. Dupé	GK	GK	['GK']
3	17374	I. Dahlqvist	RM	RM	['RM']
5	1838	M. Eikrem	CAM	CAM	['CAM']
6	9050	Francés	CB	CB	['CB']
...
5672	10888	R. Takao	CB	CB	['RB', 'CB']
5673	17487	T. Tattermusch	ST	ST	['ST']
5674	14857	Hong Jin Gi	CB	CB	['CB', 'LB']
5675	7222	E. López	CAM	CAM	['CAM', 'RM']
5676	2675	M. El Shenawy	GK	GK	['GK']



Ngoài ra, do chúng ta có khá là nhiều lớp nên có thể xảy ra tình trạng mất cân bằng. Vì thế, chúng ta sẽ sử dụng thêm một hàm có tên là SMOTE cho việc cân bằng dữ liệu. Mô hình sử dụng vẫn là RandomForestClassifier với kết quả khi chạy trên tập train là: 100%, trên tập test là: 70% mà không có gì thay đổi so với khi không sử dụng SMOTE. Hình bên trái là confusion matrix đã chuẩn hóa sang % cho 15 lớp khi chạy trên tập test có sử dụng SMOTE

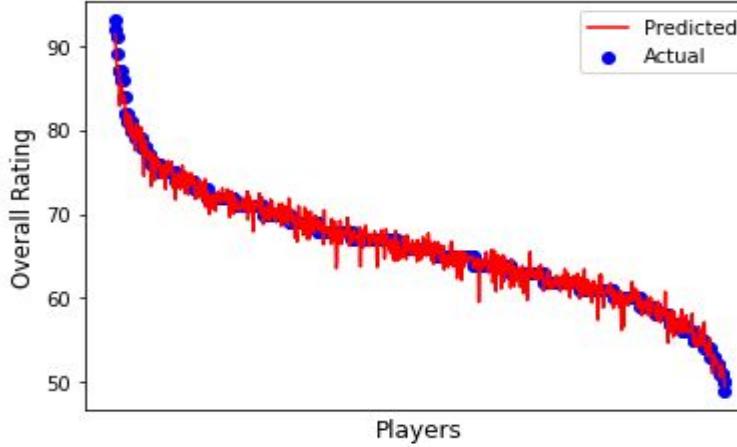
		LS	ST	RS
		89+3	89+3	89+3
LW	LF	CF	RF	RW
92+0	93+0	93+0	93+0	92+0
LAM	CAM	RAM		
93+0	93+0	93+0		
LM	LCM	CM	RCM	RM
91+2	87+3	87+3	87+3	91+2
LWB	LDM	CDM	RDM	RWB
66+3	65+3	65+3	65+3	66+3
LB	LCB	CB	RCB	RB
62+3	52+3	52+3	52+3	62+3
		GK		
		19+3		

Ở câu hỏi thứ hai, chúng ta cũng sẽ chia thành hai cách là dự đoán chỉ số tổng quát cho từng cầu thủ chỉ thuộc duy nhất một trong ba nhóm là: xanh, đỏ hoặc vàng và dự đoán chỉ số tổng quát cho từng cầu thủ thuộc bất kỳ nhóm nào. Vì vị trí thủ môn có các thuộc tính khác biệt so với các vị trí khác nên ở cả hai cách các hàng mà có vị trí thủ môn sẽ bị xóa đi

Ở cách môt, mô hình mà chúng ta sử dụng là linear regression với tỷ lệ train/test là: 7.5/2.5. Độ chính xác khi chạy trên tập train là: 98% và trên tập test là: 98%. Dưới đây là hình minh họa kết quả dự đoán của mô hình cho nhóm các cầu thủ chỉ thuộc vị trí tấn công

	short_name	Overall Rating	Predicted Overall Rating	Difference
0	L. Messi	93	91	-2.150538
1	R. Lewandowski	92	89	-3.260870
5	Neymar Jr	91	87	-4.395604
13	H. Kane	89	87	-2.247191
41	L. Suárez	87	87	0.000000
...
18591	Jeong Han Min	51	52	1.960784
18633	A. Caia	51	52	1.960784
18691	H. Ishola	50	51	2.000000
18724	A. Ferguson	50	51	2.000000
18826	Gao Xiang	49	49	0.000000

Dưới đây là biểu đồ thể hiện kết quả thực tế so với kết quả khi dự đoán bằng mô hình



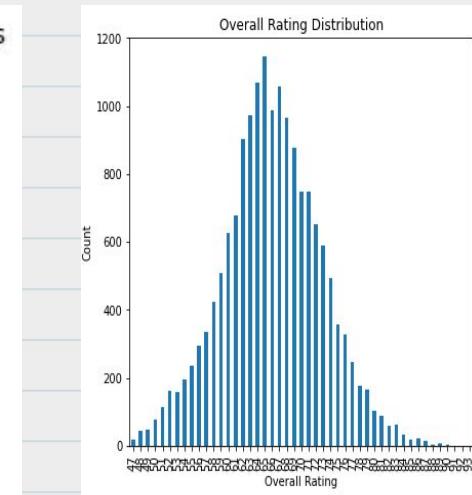
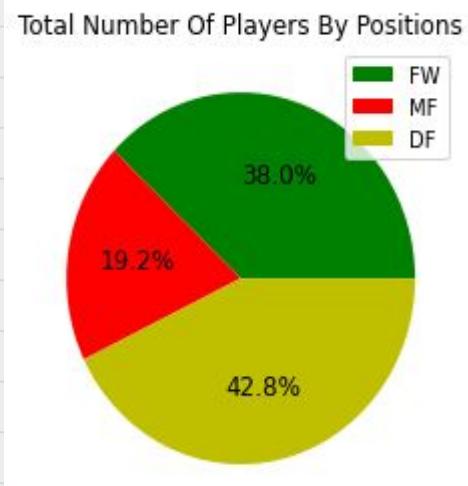
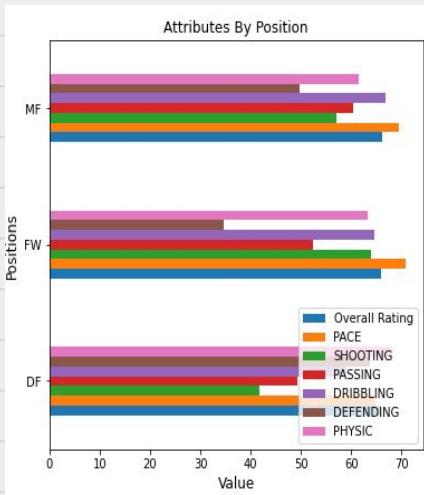


Ở cách hai, thay vì sử dụng trực tiếp các thuộc tính của mỗi cầu thủ để dự đoán chỉ số tổng quát cho toàn bộ các nhóm, chúng ta sẽ gộp các thuộc tính lại thành sáu nhóm như hình bên phải bao gồm: Pace, Shooting, Passing, Dribbling, Defending, Physic và tính trung bình của các nhóm sau khi đã gộp

Bên cạnh sáu cột này, chúng ta sẽ sử dụng thêm bốn cột khác là: 'Potential', 'age', 'Value' và 'primary_position'. Mô hình sử dụng là linear regression với tỉ lệ train/test là: 9/1. Độ chính xác khi chạy trên tập train là: 90%, trên tập test là: 89%. Dưới đây là hình minh họa kết quả dự đoán của mô hình với tổng các hàng dự đoán đúng là: 326/1683

	Order	Name	Overall Rating	Predicted Overall Rating	Difference
0	5553	V. Černý	69	66	-4.347826
1	5562	P. Klement	69	67	-2.898551
2	3434	Guto Milazar	72	69	-4.166667
3	12704	K. Kurokawa	63	62	-1.587302
4	10027	S. Ylätupa	65	64	-1.538462

Dưới đây là các hình minh họa chỉ số tổng quát, trung bình các thuộc tính của các cầu thủ theo từng vị trí; tổng số lượng cầu thủ ở các vị trí và sự phân phối chỉ số tổng quát



Link tham khảo

- [1]: Scrapy 2.5 documentation. Scrapy 2.5 documentation - Scrapy 2.5.0 documentation. (2021, April 7). Retrieved September 9, 2021, from <https://docs.scrapy.org/en/latest/>.
- [2]: Borjigin, K. (2021, September 2). Players. SoFIFA. Retrieved September 9, 2021, from <https://sofifa.com/players?col=oa&sort=desc&offset=0>.
- [3]: Wikimedia Foundation. (2021, September 5). FIFA 21. Wikipedia. Retrieved September 9, 2021, from https://en.wikipedia.org/wiki/FIFA_21.
- [4]: Brownlee, J. (2021, March 16). Smote for imbalanced classification with python. Machine Learning Mastery. Retrieved September 9, 2021, from <https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/>.

Do you have
any questions?

