In [3]:
```
import pandas as pd
```

In [4]:
```
df=pd.read_csv("fifa21 raw data v2.csv")
```

```
C:\Users\yashl\AppData\Local\Temp\ipykernel_6048\1850234331.py:1: DtypeWarning: Columns (76) have mixed types. Specify dtype option on import or set low_memory=False.
  df=pd.read_csv("fifa21 raw data v2.csv")
```

UNDERSTANDING THE DATA

In [5]:
```
df
```

Out[5]:

	ID	Name	LongName	photoUrl	playerUrl	Nationality	Age	↓OVA	POT	Club	...	A/W	D/W	IR	PAC	SHO	PAS	DRI	DEF	PHY	Hits
0	158023	L. Messi	Lionel Messi	https://cdn.sofifa.com/players/158/023/21_60.png	http://sofifa.com/player/158023/lionel-messi/2...	Argentina	33	93	93	\n\n\n\nFC Barcelona	...	Medium	Low	5 ★	85	92	91	95	38	65	771
1	20801	Cristiano Ronaldo	C. Ronaldo dos Santos Aveiro	https://cdn.sofifa.com/players/020/801/21_60.png	http://sofifa.com/player/20801/c-ronaldo-dos-s...	Portugal	35	92	92	\n\n\n\nJuventus	...	High	Low	5 ★	89	93	81	89	35	77	562
2	200389	J. Oblak	Jan Oblak	https://cdn.sofifa.com/players/200/389/21_60.png	http://sofifa.com/player/200389/jan-oblak/210006/	Slovenia	27	91	93	\n\n\n\nAtlético Madrid	...	Medium	Medium	3 ★	87	92	78	90	52	90	150
3	192985	K. De Bruyne	Kevin De Bruyne	https://cdn.sofifa.com/players/192/985/21_60.png	http://sofifa.com/player/192985/kevin-de-bruyn...	Belgium	29	91	91	\n\n\n\nManchester City	...	High	High	4 ★	76	86	93	88	64	78	207
4	190871	Neymar Jr	Neymar da Silva Santos Jr.	https://cdn.sofifa.com/players/190/871/21_60.png	http://sofifa.com/player/190871/neymar-da-silv...	Brazil	28	91	91	\n\n\n\nParis Saint-Germain	...	High	Medium	5 ★	91	85	86	94	36	59	595
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
18974	247223	Xia Ao	Ao Xia	https://cdn.sofifa.com/players/247/223/21_60.png	http://sofifa.com/player/247223/ao-xia/210006/	China PR	21	47	55	\n\n\n\nWuhan Zall	...	Medium	Medium	1 ★	64	28	26	38	48	51	NaN
18975	258760	B. Hough	Ben Hough	https://cdn.sofifa.com/players/258/760/21_60.png	http://sofifa.com/player/258760/ben-hough/210006/	England	17	47	67	\n\n\n\nOldham Athletic	...	Medium	Medium	1 ★	64	40	48	49	35	45	NaN
18976	252757	R. McKinley	Ronan McKinley	https://cdn.sofifa.com/players/252/757/21_60.png	http://sofifa.com/player/252757/ronan-mckinley...	England	18	47	65	\n\n\n\nDerry City	...	Medium	Medium	1 ★	63	39	44	46	40	53	NaN
18977	243790	Wang Zhen'ao	Zhen'ao Wang	https://cdn.sofifa.com/players/243/790/21_60.png	http://sofifa.com/player/243790/zhenao-wang/21...	China PR	20	47	57	\n\n\n\nDalian YiFang FC	...	Medium	Medium	1 ★	58	49	41	49	30	44	NaN
18978	252520	Zhou Xiao	Xiao Zhou	https://cdn.sofifa.com/players/252/520/21_60.png	http://sofifa.com/player/252520/xiao-zhou/210006/	China PR	21	47	57	\n\n\n\nDalian YiFang FC	...	Medium	Medium	1 ★	62	22	39	42	45	55	NaN

18979 rows × 77 columns

In [6]:
```
df.shape
```

Out[6]:
```
(18979, 77)
```

In [7]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 18979 entries, 0 to 18978
Data columns (total 77 columns):
 #   Column            Non-Null Count  Dtype 
---  ------            --------------  ----- 
 0   ID                18979 non-null  int64 
 1   Name              18979 non-null  object
 2   LongName          18979 non-null  object
 3   photoUrl          18979 non-null  object
 4   playerUrl         18979 non-null  object
 5   Nationality       18979 non-null  object
 6   Age               18979 non-null  int64 
 7   ↓OVA              18979 non-null  int64 
 8   POT               18979 non-null  int64 
 9   Club              18979 non-null  object
 10  Contract          18979 non-null  object
 11  Positions         18979 non-null  object
 12  Height            18979 non-null  object
 13  Weight            18979 non-null  object
 14  Preferred Foot    18979 non-null  object
 15  BOV               18979 non-null  int64 
 16  Best Position     18979 non-null  object
 17  Joined            18979 non-null  object
 18  Loan Date End     1013 non-null   object
 19  Value             18979 non-null  object
 20  Wage              18979 non-null  object
 21  Release Clause    18979 non-null  object
 22  Attacking         18979 non-null  int64 
 23  Crossing          18979 non-null  int64 
 24  Finishing         18979 non-null  int64 
 25  Heading Accuracy  18979 non-null  int64 
 26  Short Passing     18979 non-null  int64 
 27  Volleys           18979 non-null  int64 
 28  Skill             18979 non-null  int64 
 29  Dribbling         18979 non-null  int64 
 30  Curve             18979 non-null  int64 
 31  FK Accuracy       18979 non-null  int64 
 32  Long Passing      18979 non-null  int64 
 33  Ball Control      18979 non-null  int64 
 34  Movement          18979 non-null  int64 
 35  Acceleration      18979 non-null  int64 
 36  Sprint Speed      18979 non-null  int64 
 37  Agility           18979 non-null  int64 
 38  Reactions         18979 non-null  int64 
 39  Balance           18979 non-null  int64 
 40  Power             18979 non-null  int64 
 41  Shot Power        18979 non-null  int64 
 42  Jumping           18979 non-null  int64 
 43  Stamina           18979 non-null  int64 
 44  Strength          18979 non-null  int64 
 45  Long Shots        18979 non-null  int64 
 46  Mentality         18979 non-null  int64 
 47  Aggression        18979 non-null  int64 
 48  Interceptions     18979 non-null  int64 
 49  Positioning       18979 non-null  int64 
 50  Vision            18979 non-null  int64 
 51  Penalties         18979 non-null  int64 
 52  Composure         18979 non-null  int64 
 53  Defending         18979 non-null  int64 
 54  Marking           18979 non-null  int64 
 55  Standing Tackle   18979 non-null  int64 
 56  Sliding Tackle    18979 non-null  int64 
 57  Goalkeeping       18979 non-null  int64 
 58  GK Diving         18979 non-null  int64 
 59  GK Handling       18979 non-null  int64 
 60  GK Kicking        18979 non-null  int64 
 61  GK Positioning    18979 non-null  int64 
 62  GK Reflexes       18979 non-null  int64 
 63  Total Stats       18979 non-null  int64 
 64  Base Stats        18979 non-null  int64 
 65  W/F               18979 non-null  object
 66  SM                18979 non-null  object
 67  A/W               18979 non-null  object
 68  D/W               18979 non-null  object
 69  IR                18979 non-null  object
 70  PAC               18979 non-null  int64 
 71  SHO               18979 non-null  int64 
 72  PAS               18979 non-null  int64 
 73  DRI               18979 non-null  int64 
 74  DEF               18979 non-null  int64 
 75  PHY               18979 non-null  int64 
 76  Hits              16384 non-null  object
dtypes: int64(54), object(23)
memory usage: 11.1+ MB
```

DATA CLEANING

In [8]:
```
df.duplicated().sum()
```

Out[8]:
```
0
```

In [9]:
```
df['ID'].duplicated().sum()
```

Out[9]:
```
0
```

In [10]:
```
df.head()
```

Out[10]:

	ID	Name	LongName	photoUrl	playerUrl	Nationality	Age	↓OVA	POT	Club	...	A/W	D/W	IR	PAC	SHO	PAS	DRI	DEF	PHY	Hits
0	158023	L. Messi	Lionel Messi	https://cdn.sofifa.com/players/158/023/21_60.png	http://sofifa.com/player/158023/lionel-messi/2...	Argentina	33	93	93	\n\n\n\nFC Barcelona	...	Medium	Low	5 ★	85	92	91	95	38	65	771
1	20801	Cristiano Ronaldo	C. Ronaldo dos Santos Aveiro	https://cdn.sofifa.com/players/020/801/21_60.png	http://sofifa.com/player/20801/c-ronaldo-dos-s...	Portugal	35	92	92	\n\n\n\nJuventus	...	High	Low	5 ★	89	93	81	89	35	77	562
2	200389	J. Oblak	Jan Oblak	https://cdn.sofifa.com/players/200/389/21_60.png	http://sofifa.com/player/200389/jan-oblak/210006/	Slovenia	27	91	93	\n\n\n\nAtlético Madrid	...	Medium	Medium	3 ★	87	92	78	90	52	90	150
3	192985	K. De Bruyne	Kevin De Bruyne	https://cdn.sofifa.com/players/192/985/21_60.png	http://sofifa.com/player/192985/kevin-de-bruyn...	Belgium	29	91	91	\n\n\n\nManchester City	...	High	High	4 ★	76	86	93	88	64	78	207
4	190871	Neymar Jr	Neymar da Silva Santos Jr.	https://cdn.sofifa.com/players/190/871/21_60.png	http://sofifa.com/player/190871/neymar-da-silv...	Brazil	28	91	91	\n\n\n\nParis Saint-Germain	...	High	Medium	5 ★	91	85	86	94	36	59	595

5 rows × 77 columns

REMOVE UNNECESSARY COLUMNS

In [11]:
```
df = df.drop(["LongName","photoUrl","playerUrl"],axis=1)
```

In [12]:
```
df
```

Out[12]:

	ID	Name	Nationality	Age	↓OVA	POT	Club	Contract	Positions	Height	...	A/W	D/W	IR	PAC	SHO	PAS	DRI	DEF	PHY	Hits
0	158023	L. Messi	Argentina	33	93	93	\n\n\n\nFC Barcelona	2004 ~ 2021	RW, ST, CF	170cm	...	Medium	Low	5 ★	85	92	91	95	38	65	771
1	20801	Cristiano Ronaldo	Portugal	35	92	92	\n\n\n\nJuventus	2018 ~ 2022	ST, LW	187cm	...	High	Low	5 ★	89	93	81	89	35	77	562
2	200389	J. Oblak	Slovenia	27	91	93	\n\n\n\nAtlético Madrid	2014 ~ 2023	GK	188cm	...	Medium	Medium	3 ★	87	92	78	90	52	90	150
3	192985	K. De Bruyne	Belgium	29	91	91	\n\n\n\nManchester City	2015 ~ 2023	CAM, CM	181cm	...	High	High	4 ★	76	86	93	88	64	78	207
4	190871	Neymar Jr	Brazil	28	91	91	\n\n\n\nParis Saint-Germain	2017 ~ 2022	LW, CAM	175cm	...	High	Medium	5 ★	91	85	86	94	36	59	595
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
18974	247223	Xia Ao	China PR	21	47	55	\n\n\n\nWuhan Zall	2018 ~ 2022	CB	178cm	...	Medium	Medium	1 ★	64	28	26	38	48	51	NaN
18975	258760	B. Hough	England	17	47	67	\n\n\n\nOldham Athletic	2020 ~ 2021	CM	175cm	...	Medium	Medium	1 ★	64	40	48	49	35	45	NaN
18976	252757	R. McKinley	England	18	47	65	\n\n\n\nDerry City	2019 ~ 2020	CM	179cm	...	Medium	Medium	1 ★	63	39	44	46	40	53	NaN
18977	243790	Wang Zhen'ao	China PR	20	47	57	\n\n\n\nDalian YiFang FC	2020 ~ 2022	RW	175cm	...	Medium	Medium	1 ★	58	49	41	49	30	44	NaN
18978	252520	Zhou Xiao	China PR	21	47	57	\n\n\n\nDalian YiFang FC	2019 ~ 2023	CB, LB	188cm	...	Medium	Medium	1 ★	62	22	39	42	45	55	NaN

18979 rows × 74 columns

RENAME COLUMNS SO THEY ARE OF THE SAME TYPE

In [13]:
```
df=df.rename(columns={"↓OVA":"OVA"})
```

In [14]:
```
df
```

Out[14]:

	ID	Name	Nationality	Age	OVA	POT	Club	Contract	Positions	Height	...	A/W	D/W	IR	PAC	SHO	PAS	DRI	DEF	PHY	Hits
0	158023	L. Messi	Argentina	33	93	93	\n\n\n\nFC Barcelona	2004 ~ 2021	RW, ST, CF	170cm	...	Medium	Low	5 ★	85	92	91	95	38	65	771
1	20801	Cristiano Ronaldo	Portugal	35	92	92	\n\n\n\nJuventus	2018 ~ 2022	ST, LW	187cm	...	High	Low	5 ★	89	93	81	89	35	77	562
2	200389	J. Oblak	Slovenia	27	91	93	\n\n\n\nAtlético Madrid	2014 ~ 2023	GK	188cm	...	Medium	Medium	3 ★	87	92	78	90	52	90	150
3	192985	K. De Bruyne	Belgium	29	91	91	\n\n\n\nManchester City	2015 ~ 2023	CAM, CM	181cm	...	High	High	4 ★	76	86	93	88	64	78	207
4	190871	Neymar Jr	Brazil	28	91	91	\n\n\n\nParis Saint-Germain	2017 ~ 2022	LW, CAM	175cm	...	High	Medium	5 ★	91	85	86	94	36	59	595
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
18974	247223	Xia Ao	China PR	21	47	55	\n\n\n\nWuhan Zall	2018 ~ 2022	CB	178cm	...	Medium	Medium	1 ★	64	28	26	38	48	51	NaN
18975	258760	B. Hough	England	17	47	67	\n\n\n\nOldham Athletic	2020 ~ 2021	CM	175cm	...	Medium	Medium	1 ★	64	40	48	49	35	45	NaN
18976	252757	R. McKinley	England	18	47	65	\n\n\n\nDerry City	2019 ~ 2020	CM	179cm	...	Medium	Medium	1 ★	63	39	44	46	40	53	NaN
18977	243790	Wang Zhen'ao	China PR	20	47	57	\n\n\n\nDalian YiFang FC	2020 ~ 2022	RW	175cm	...	Medium	Medium	1 ★	58	49	41	49	30	44	NaN
18978	252520	Zhou Xiao	China PR	21	47	57	\n\n\n\nDalian YiFang FC	2019 ~ 2023	CB, LB	188cm	...	Medium	Medium	1 ★	62	22	39	42	45	55	NaN

18979 rows × 74 columns

REMOVE UNNECESSARY NEWLINE CHARACTERS

In [15]:
```
df['Club']=df['Club'].str.strip('\n')
```

In [16]:
```
df
```

Out[16]:

	ID	Name	Nationality	Age	OVA	POT	Club	Contract	Positions	Height	...	A/W	D/W	IR	PAC	SHO	PAS	DRI	DEF	PHY	Hits
0	158023	L. Messi	Argentina	33	93	93	FC Barcelona	2004 ~ 2021	RW, ST, CF	170cm	...	Medium	Low	5 ★	85	92	91	95	38	65	771
1	20801	Cristiano Ronaldo	Portugal	35	92	92	Juventus	2018 ~ 2022	ST, LW	187cm	...	High	Low	5 ★	89	93	81	89	35	77	562
2	200389	J. Oblak	Slovenia	27	91	93	Atlético Madrid	2014 ~ 2023	GK	188cm	...	Medium	Medium	3 ★	87	92	78	90	52	90	150
3	192985	K. De Bruyne	Belgium	29	91	91	Manchester City	2015 ~ 2023	CAM, CM	181cm	...	High	High	4 ★	76	86	93	88	64	78	207
4	190871	Neymar Jr	Brazil	28	91	91	Paris Saint-Germain	2017 ~ 2022	LW, CAM	175cm	...	High	Medium	5 ★	91	85	86	94	36	59	595
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
18974	247223	Xia Ao	China PR	21	47	55	Wuhan Zall	2018 ~ 2022	CB	178cm	...	Medium	Medium	1 ★	64	28	26	38	48	51	NaN
18975	258760	B. Hough	England	17	47	67	Oldham Athletic	2020 ~ 2021	CM	175cm	...	Medium	Medium	1 ★	64	40	48	49	35	45	NaN
18976	252757	R. McKinley	England	18	47	65	Derry City	2019 ~ 2020	CM	179cm	...	Medium	Medium	1 ★	63	39	44	46	40	53	NaN
18977	243790	Wang Zhen'ao	China PR	20	47	57	Dalian YiFang FC	2020 ~ 2022	RW	175cm	...	Medium	Medium	1 ★	58	49	41	49	30	44	NaN
18978	252520	Zhou Xiao	China PR	21	47	57	Dalian YiFang FC	2019 ~ 2023	CB, LB	188cm	...	Medium	Medium	1 ★	62	22	39	42	45	55	NaN

18979 rows × 74 columns

CONVERTING HEIGHT AND WEIGHT TO NUMERIC FORM

In [17]:
```
df["Height"]=df['Height'].str.strip('cm')
```

In [18]:
```
df["Height"].unique()
```

Out[18]:
```
array(['170', '187', '188', '181', '175', '184', '191', '178', '193',
       '185', '199', '173', '168', '176', '177', '183', '180', '189',
       '179', '195', '172', '182', '186', '192', '165', '194', '167',
       '196', '163', '190', '174', '169', '171', '197', '200', '166',
       '6\'2"', '164', '198', '6\'3"', '6\'5"', '5\'11"', '6\'4"',
       '6\'1"', '6\'0"', '5\'10"', '5\'9"', '5\'6"', '5\'7"', '5\'4"',
       '201', '158', '162', '161', '160', '203', '157', '156', '202',
       '159', '206', '155'], dtype=object)
```

In [19]:
```
def ft_to_cm(x):
    if "'" in x:
        ft=x.replace("'",".").replace('"',"")
        cm=round(float(ft)*30.48,0)
        return int(cm)
    else:
        return int(x)

df["Height"]=df["Height"].apply(ft_to_cm)
df["Height"].unique()
```

Out[19]:
```
array([170, 187, 188, 181, 175, 184, 191, 178, 193, 185, 199, 173, 168,
       176, 177, 183, 180, 189, 179, 195, 172, 182, 186, 192, 165, 194,
       167, 196, 163, 190, 174, 169, 171, 197, 200, 166, 198, 156, 155,
       155, 201, 158, 162, 161, 160, 203, 157, 202, 159, 206], dtype=int64)
```

In [20]:
```
df.columns
```

Out[20]:
```
Index(['ID', 'Name', 'Nationality', 'Age', 'OVA', 'POT', 'Club', 'Contract',
       'Positions', 'Height', 'Weight', 'Preferred Foot', 'BOV',
       'Best Position', 'Joined', 'Loan Date End', 'Value', 'Wage',
       'Release Clause', 'Attacking', 'Crossing', 'Finishing',
       'Heading Accuracy', 'Short Passing', 'Volleys', 'Skill', 'Dribbling',
       'Curve', 'FK Accuracy', 'Long Passing', 'Ball Control', 'Movement',
       'Acceleration', 'Sprint Speed', 'Agility', 'Reactions', 'Balance',
       'Power', 'Shot Power', 'Jumping', 'Stamina', 'Strength', 'Long Shots',
       'Mentality', 'Aggression', 'Interceptions', 'Positioning', 'Vision',
       'Penalties', 'Composure', 'Defending', 'Marking', 'Standing Tackle',
       'Sliding Tackle', 'Goalkeeping', 'GK Diving', 'GK Handling',
       'GK Kicking', 'GK Positioning', 'GK Reflexes', 'Total Stats',
       'Base Stats', 'W/F', 'SM', 'A/W', 'D/W', 'IR', 'PAC', 'SHO', 'PAS',
       'DRI', 'DEF', 'PHY', 'Hits'],
      dtype='object')
```

In [21]:
```
df["Weight"]
```

Out[21]:
```
0        72kg
1        83kg
2        87kg
3        70kg
4        68kg
         ... 
18974    66kg
18975    65kg
18976    74kg
18977    69kg
18978    75kg
Name: Weight, Length: 18979, dtype: object
```

In [22]:
```
df["Weight"]=df["Weight"].str.strip("kg")
```

In [23]:
```
df["Weight"].unique()
```

Out[23]:
```
array(['72', '83', '87', '70', '68', '80', '71', '91', '73', '85', '92',
       '69', '84', '96', '81', '82', '75', '86', '89', '74', '76', '64',
       '78', '90', '66', '60', '94', '79', '67', '65', '59', '61', '93',
       '88', '97', '77', '62', '63', '95', '100', '58', '183lbs',
       '179lbs', '172lbs', '196lbs', '176lbs', '185lbs', '170lbs',
       '203lbs', '168lbs', '161lbs', '146lbs', '130lbs', '190lbs',
       '174lbs', '148lbs', '165lbs', '159lbs', '192lbs', '181lbs',
       '139lbs', '154lbs', '157lbs', '163lbs', '98', '103', '99', '102',
       '56', '101', '57', '55', '104', '107', '110', '53', '50', '54',
       '52'], dtype=object)
```

In [24]:
```
def lbs_to_kg(x):
    if "lbs" in x:
        lbs = x.replace("lbs","")
        kg = round(int(lbs)*0.453,0)
        return int(kg)
    else:
        return int(x)
df["Weight"]=df["Weight"].apply(lbs_to_kg)
```

In [25]:
```
df["Weight"].unique()
```

Out[25]:
```
array([ 72,  83,  87,  70,  68,  80,  71,  91,  73,  85,  92,  69,  84,
        96,  81,  82,  75,  86,  89,  74,  76,  64,  78,  90,  66,  60,
        94,  79,  67,  65,  59,  61,  93,  88,  97,  77,  62,  63,  95,
       100,  58,  98, 103,  99, 102,  56, 101,  57,  55, 104, 107, 110,
        53,  50,  54,  52], dtype=int64)
```

In [26]:
```
df.columns
```

Out[26]:
```
Index(['ID', 'Name', 'Nationality', 'Age', 'OVA', 'POT', 'Club', 'Contract',
       'Positions', 'Height', 'Weight', 'Preferred Foot', 'BOV',
       'Best Position', 'Joined', 'Loan Date End', 'Value', 'Wage',
       'Release Clause', 'Attacking', 'Crossing', 'Finishing',
       'Heading Accuracy', 'Short Passing', 'Volleys', 'Skill', 'Dribbling',
       'Curve', 'FK Accuracy', 'Long Passing', 'Ball Control', 'Movement',
       'Acceleration', 'Sprint Speed', 'Agility', 'Reactions', 'Balance',
       'Power', 'Shot Power', 'Jumping', 'Stamina', 'Strength', 'Long Shots',
       'Mentality', 'Aggression', 'Interceptions', 'Positioning', 'Vision',
       'Penalties', 'Composure', 'Defending', 'Marking', 'Standing Tackle',
       'Sliding Tackle', 'Goalkeeping', 'GK Diving', 'GK Handling',
       'GK Kicking', 'GK Positioning', 'GK Reflexes', 'Total Stats',
       'Base Stats', 'W/F', 'SM', 'A/W', 'D/W', 'IR', 'PAC', 'SHO', 'PAS',
       'DRI', 'DEF', 'PHY', 'Hits'],
      dtype='object')
```

CONVERT THE STRING COLUMNS TO NUMERIC

In [27]:
```
df["Value"].unique()
```

Out[27]:
```
array(['€103.5M', '€63M', '€120M', '€129M', '€132M', '€111M', '€120.5M',
       '€102M', '€185.5M', '€110M', '€113M', '€90.5M', '€82M', '€17.5M',
       '€83.5M', '€33.5M', '€114.5M', '€78M', '€103M', '€109M', '€92M',
       '€10M', '€76.5M', '€89.5M', '€87.5M', '€79.5M', '€124M', '€114M',
       '€95M', '€92.5M', '€105.5M', '€88.5M', '€85M', '€81.5M', '€26M',
       '€21M', '€56M', '€67.5M', '€53M', '€36.5M', '€51M', '€65.5M',
       '€46.5M', '€61.5M', '€72.5M', '€77.5M', '€43.5M', '€32.5M', '€36M',
       '€32M', '€54M', '€49.5M', '€57M', '€66.5M', '€74.5M', '€71.5M',
       '€121M', '€99M', '€67M', '€86.5M', '€93.5M', '€70M', '€62M',
       '€66M', '€58M', '€44M', '€81M', '€37M', '€14.5M', '€46M', '€47.5M',
       '€52.5M', '€54.5M', '€34.5M', '€57.5M', '€51.5M', '€44.5M', '€55M',
       '€48M', '€60.5M', '€63.5M', '€61M', '€29M', '€58.5M', '€55.5M',
       '€42M', '€40.5M', '€43M', '€45.5M', '€34M', '€26.5M', '€42.5M',
       '€35.5M', '€45M', '€41.5M', '€40M', '€11M', '€13.5M', '€29.5M',
       '€27M', '€15.5M', '€38.5M', '€52M', '€33M', '€19M', '€73.5M',
       '€38M', '€35M', '€47M', '€24M', '€30.5M', '€18M', '€28M', '€25.5M',
       '€25M', '€31M', '€23.5M', '€30M', '€31.5M', '€22.5M', '€28.5M',
       '€4M', '€12.5M', '€37.5M', '€27.5M', '€16M', '€15M', '€20.5M',
       '€22M', '€3.4M', '€5M', '€56.5M', '€62.5M', '€0', '€39M', '€24.5M',
       '€21.5M', '€13M', '€8M', '€20M', '€8.5M', '€2.9M', '€9M', '€4.6M',
       '€50M', '€23M', '€18.5M', '€7M', '€19.5M', '€5.5M', '€7.5M',
       '€3.8M', '€14M', '€10.5M', '€16.5M', '€3.6M', '€9.5M', '€39.5M',
       '€17M', '€12M', '€11.5M', '€4.9M', '€3M', '€1.9M', '€6.5M',
       '€1.7M', '€2.4M', '€3.1M', '€6M', '€3.7M', '€4.7M', '€4.3M',
       '€2.1M', '€1.2M', '€1.8M', '€4.8M', '€3.2M', '€1.3M', '€825K',
       '€2.3M', '€1.5M', '€3.9M', '€2.6M', '€3.5M', '€2.8M', '€2.7M',
       '€4.4M', '€4.1M', '€950K', '€1.6M', '€625K', '€1.1M', '€4.5M',
       '€4.2M', '€2.2M', '€3.3M', '€1.4M', '€2M', '€475K', '€925K',
       '€750K', '€725K', '€2.5M', '€1M', '€350K', '€525K', '€600K',
       '€850K', '€800K', '€550K', '€250K', '€400K', '€425K', '€575K',
       '€210K', '€325K', '€900K', '€875K', '€650K', '€700K', '€500K',
       '€975K', '€375K', '€775K', '€275K', '€180K', '€450K', '€675K',
       '€150K', '€240K', '€300K', '€130K', '€220K', '€200K', '€110K',
       '€170K', '€230K', '€90K', '€120K', '€80K', '€190K', '€140K',
       '€160K', '€100K', '€60K', '€50K', '€70K', '€45K', '€35K', '€40K',
       '€25K', '€20K', '€15K', '€30K', '€9K'], dtype=object)
```

In [28]:
```
def money(x):
    if "€" in x:
        x=x.strip("€")
    if "M" in x:
        x=x.strip("M")
        x=float(x)*1000000
        return int(x)
    elif "K" in x:
        x=x.strip("K")
        x=float(x)*1000
        return int(x)
    else:
        return int(x)

df["Value"]=df["Value"].apply(money)
```

In [29]:
```
df["Wage"]=df["Wage"].apply(money)
```

In [30]:
```
df["Release Clause"]=df["Release Clause"].apply(money)
```

In [31]:
```
df.columns
```

Out[31]:
```
Index(['ID', 'Name', 'Nationality', 'Age', 'OVA', 'POT', 'Club', 'Contract',
       'Positions', 'Height', 'Weight', 'Preferred Foot', 'BOV',
       'Best Position', 'Joined', 'Loan Date End', 'Value', 'Wage',
       'Release Clause', 'Attacking', 'Crossing', 'Finishing',
       'Heading Accuracy', 'Short Passing', 'Volleys', 'Skill', 'Dribbling',
       'Curve', 'FK Accuracy', 'Long Passing', 'Ball Control', 'Movement',
       'Acceleration', 'Sprint Speed', 'Agility', 'Reactions', 'Balance',
       'Power', 'Shot Power', 'Jumping', 'Stamina', 'Strength', 'Long Shots',
       'Mentality', 'Aggression', 'Interceptions', 'Positioning', 'Vision',
       'Penalties', 'Composure', 'Defending', 'Marking', 'Standing Tackle',
       'Sliding Tackle', 'Goalkeeping', 'GK Diving', 'GK Handling',
       'GK Kicking', 'GK Positioning', 'GK Reflexes', 'Total Stats',
       'Base Stats', 'W/F', 'SM', 'A/W', 'D/W', 'IR', 'PAC', 'SHO', 'PAS',
       'DRI', 'DEF', 'PHY', 'Hits'],
      dtype='object')
```

REMOVE THE STARS FROM VALUES

In [32]:
```
df["W/F"].unique()
```

Out[32]:
```
array(['4 ★', '3 ★', '5 ★', '2 ★', '1 ★'], dtype=object)
```

In [33]:
```
df["W/F"]=df["W/F"].str.replace("★","")
```

In [34]:
```
df["SM"]=df["SM"].str.replace("★","")
```

In [35]:
```
df["IR"]=df["IR"].str.replace("★","")
```

CLEANING UP THE CONTRACT COLUMN

In [36]:
```
df["Contract"].unique()
```

Out[36]:
```
array(['2004 ~ 2021', '2018 ~ 2022', '2014 ~ 2023', '2015 ~ 2023',
       '2017 ~ 2022', '2017 ~ 2023', '2018 ~ 2024', '2014 ~ 2022',
       '2018 ~ 2023', '2016 ~ 2023', '2013 ~ 2023', '2011 ~ 2023',
       '2009 ~ 2022', '2005 ~ 2021', '2011 ~ 2021', '2015 ~ 2022',
       '2017 ~ 2024', '2010 ~ 2024', '2012 ~ 2021', '2019 ~ 2024',
       '2015 ~ 2024', '2017 ~ 2025', '2020 ~ 2025', '2019 ~ 2023',
       '2008 ~ 2023', '2015 ~ 2021', '2020 ~ 2022', '2012 ~ 2022',
       '2016 ~ 2025', '2013 ~ 2022', '2011 ~ 2022', '2012 ~ 2024',
       '2016 ~ 2021', '2012 ~ 2023', '2008 ~ 2022', '2019 ~ 2022',
       '2017 ~ 2021', '2013 ~ 2024', '2020 ~ 2024', '2010 ~ 2022',
       '2020 ~ 2021', '2011 ~ 2024', '2020 ~ 2023', '2014 ~ 2024',
       '2013 ~ 2026', '2016 ~ 2022', '2010 ~ 2021', '2013 ~ 2021',
       '2019 ~ 2025', '2018 ~ 2025', '2016 ~ 2024', '2018 ~ 2021',
       '2009 ~ 2024', '2007 ~ 2022', 'Jun 30, 2021 On Loan',
       '2009 ~ 2021', '2019 ~ 2021', '2019 ~ 2026', 'Free', '2012 ~ 2028',
       '2010 ~ 2023', '2014 ~ 2021', '2015 ~ 2025', '2014 ~ 2026',
       '2012 ~ 2025', '2017 ~ 2020', '2002 ~ 2022', '2020 ~ 2027',
       '2013 ~ 2025', 'Dec 31, 2020 On Loan', '2019 ~ 2020',
       '2011 ~ 2025', '2016 ~ 2020', '2007 ~ 2021', '2020 ~ 2026',
       '2010 ~ 2025', '2009 ~ 2023', '2008 ~ 2021', '2020 ~ 2020',
       '2016 ~ 2026', 'Jan 30, 2021 On Loan', '2012 ~ 2020',
       '2014 ~ 2025', 'Jun 30, 2022 On Loan', '2015 ~ 2020',
       'May 31, 2021 On Loan', '2018 ~ 2020', '2014 ~ 2020',
       '2013 ~ 2020', '2006 ~ 2024', 'Jul 5, 2021 On Loan',
       'Dec 31, 2021 On Loan', '2004 ~ 2025', '2011 ~ 2020',
       'Jul 1, 2021 On Loan', 'Jan 1, 2021 On Loan', '2006 ~ 2023',
       'Aug 31, 2021 On Loan', '2006 ~ 2021', '2005 ~ 2023',
       '2003 ~ 2020', '2009 ~ 2020', '2002 ~ 2020', '2005 ~ 2020',
       '2005 ~ 2022', 'Jan 31, 2021 On Loan', '2010 ~ 2020',
       'Dec 30, 2021 On Loan', '2008 ~ 2020', '2007 ~ 2020',
       '2003 ~ 2021', 'Jun 23, 2021 On Loan', 'Jan 3, 2021 On Loan',
       'Nov 27, 2021 On Loan', '2002 ~ 2021', 'Jan 17, 2021 On Loan',
       'Jun 30, 2023 On Loan', '1998 ~ 2021', '2003 ~ 2022',
       '2007 ~ 2023', 'Jul 31, 2021 On Loan', 'Nov 22, 2020 On Loan',
       'May 31, 2022 On Loan', '2006 ~ 2020', 'Dec 30, 2020 On Loan',
       '2007 ~ 2025', 'Jan 4, 2021 On Loan', 'Nov 30, 2020 On Loan',
       '2004 ~ 2020', '2009 ~ 2025', 'Aug 1, 2021 On Loan'], dtype=object)
```

In [37]:
```
from datetime import datetime, timedelta
```

In [38]:
```
def type(x):
    if "Free" in x:
        return "Free"
    if "Loan" in x:
        return "Loan"
    if "~" in x:
        return "Contract"
    else:
        return pd.NA


def start_time_contract(x):
    if "~" in x:
        return  int(x[:4])
    if "Loan" in x:
        x=x.strip(" On Loan")
        x= datetime.strptime(x, "%b %d, %Y")
        return x.date()
    else:
        return pd.NA

def end_time_contract(type,contract,loan):
    if type == "Contract":
        return int(contract[-4:])
    if type == "Loan":
        loan = datetime.strptime(loan, "%b %d, %Y")
        return loan.date()

    else:
        return pd.NA
```

In [39]:
```
df["Type of contract"] = df["Contract"].apply(type)
```

In [40]:
```
df["Start year"] = df["Contract"].apply(start_time_contract)
```

In [41]:
```
df["End year"] = df.apply(lambda row: end_time_contract(row["Type of contract"],row["Contract"],row["Loan Date End"] ), axis=1)
```

In [42]:
```
df=df[list(df.columns[:7])+list(df.columns[-3:])+list(df.columns[8:-3])]
df=df.drop(columns=["Loan Date End"])
```

In [43]:
```
df["Attacking"]=round(df["Attacking"]/5,0)
df["Attacking"]=df["Attacking"].astype("int64")
```

In [44]:
```
df["Skill"]=df["Skill"]/5
df["Skill"]=df["Skill"].astype("int64")
```

In [45]:
```
df["Movement"]=df["Movement"]/5
df["Movement"]=df["Movement"].astype("int64")
```

In [46]:
```
df["Power"]=df["Power"]/5
df["Power"]=df["Power"].astype("int64")
```

In [47]:
```
df["Mentality"]=df["Mentality"]/6
df["Mentality"]=df["Mentality"].astype("int64")

df["Defending"]=df["Defending"]/3
df["Defending"]=df["Defending"].astype("int64")

df["Goalkeeping"] = df["Goalkeeping"]/5
df["Goalkeeping"] = df["Goalkeeping"].astype("int64")
```

In [48]:
```
df=df.rename(columns={"Attacking":"Attacking AVG", "Skill":"Skill AVG", "Movement":"Movement AVG",
                      "Power":"Power AVG", "Mentality":"Mentality AVG", "Defending":"Defending AVG", "Goalkeeping":"Goalkeeping AVG"} )
```

CLEAN DATASET

In [49]:
```
df
```

Out[49]:

	ID	Name	Nationality	Age	OVA	POT	Club	Type of contract	Start year	End year	...	A/W	D/W	IR	PAC	SHO	PAS	DRI	DEF	PHY	Hits
0	158023	L. Messi	Argentina	33	93	93	FC Barcelona	Contract	2004	2021	...	Medium	Low	5	85	92	91	95	38	65	771
1	20801	Cristiano Ronaldo	Portugal	35	92	92	Juventus	Contract	2018	2022	...	High	Low	5	89	93	81	89	35	77	562
2	200389	J. Oblak	Slovenia	27	91	93	Atlético Madrid	Contract	2014	2023	...	Medium	Medium	3	87	92	78	90	52	90	150
3	192985	K. De Bruyne	Belgium	29	91	91	Manchester City	Contract	2015	2023	...	High	High	4	76	86	93	88	64	78	207
4	190871	Neymar Jr	Brazil	28	91	91	Paris Saint-Germain	Contract	2017	2022	...	High	Medium	5	91	85	86	94	36	59	595
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
18974	247223	Xia Ao	China PR	21	47	55	Wuhan Zall	Contract	2018	2022	...	Medium	Medium	1	64	28	26	38	48	51	NaN
18975	258760	B. Hough	England	17	47	67	Oldham Athletic	Contract	2020	2021	...	Medium	Medium	1	64	40	48	49	35	45	NaN
18976	252757	R. McKinley	England	18	47	65	Derry City	Contract	2019	2020	...	Medium	Medium	1	63	39	44	46	40	53	NaN
18977	243790	Wang Zhen'ao	China PR	20	47	57	Dalian YiFang FC	Contract	2020	2022	...	Medium	Medium	1	58	49	41	49	30	44	NaN
18978	252520	Zhou Xiao	China PR	21	47	57	Dalian YiFang FC	Contract	2019	2023	...	Medium	Medium	1	62	22	39	42	45	55	NaN

18979 rows × 75 columns

In [50]:
```
df.to_csv('fifacleandata.csv',index=False)
```