Tamas Flesch Thesis - EDA step

LMJU - UpGrad - DS

Fifa 23 Ultimate Team player price prediction based on the player's attributes

Table of contents

1. Read data

Imports

In [1]:
```
# Suppressing Warnings
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
# Importing Pandas and NumPy
import pandas as pd, numpy as np
```

In [3]:
```
pd.options.display.max_columns = None
pd.options.display.max_rows = None
```

In [4]:
```
# Importing all datasets
futbin_data = pd.read_csv("fifa23_players_2023-05-30.csv")
futbin_data.head()
```

Out[4]:

	Name	Club	Nation	League	Rating	Main_Position	Alternate_Positions	Card_Version	R
0	Pelé	FUT ICONS	Brazil	Icons	98	CAM	CF,ST	Icon	
1	Lionel Messi	Paris SG	Argentina	Ligue 1	98	RW	RM	TOTY	C
2	Lionel Messi	Paris SG	Argentina	Ligue 1	98	ST	RM,RW	TOTS	C
3	Karim Benzema	Real Madrid	France	LaLiga Santander	97	CF	ST	TOTY	C
4	Kylian Mbappé	Paris SG	France	Ligue 1	97	ST	CF,LW	TOTY	C

check and understand the Price data column

In [5]:
```
futbin_data.Price.describe()
```

Out[5]:
```
count    7.160000e+03
mean     2.370740e+04
std      3.146186e+05
min      0.000000e+00
25%      2.000000e+02
50%      3.000000e+02
75%      9.000000e+02
max      1.500000e+07
Name: Price, dtype: float64
```

In [6]:
```
len(futbin_data[futbin_data.Price == 0])
```

Out[6]:
```
751
```

In [7]:
```
futbin_data[futbin_data.Price == 0].head()
```

Out[7]:

	Name	Club	Nation	League	Rating	Main_Position	Alternate_Positions	Card_Version
84	Kaoru Mitoma	Brighton	Japan	Premier League	92	LW	LWB,LM	TOTS
136	Gayà	Valencia CF	Spain	LaLiga Santander	91	LB	LWB,LM	TOTS
137	Yannick Carrasco	Atlético de Madrid	Belgium	LaLiga Santander	91	LM	LW	Flashback SBC
178	Amine Gouiri	Rennes	France	Ligue 1	90	LM	CF,ST,LW	Fantasy Objective
185	Frank Lampard	FUT ICONS	England	Icons	90	CM	CDM,CAM	Trophy Titans - ICON

SBC concept, ingame, there is a mini game, where players can finish squad building challenges (SBCs), and get valuable player. These players are not in the market, so players can't buy them on the market, that's why they havn't got a price. So I'll remove these cards from the dataset.

In [8]:
```
futbin_data = futbin_data[futbin_data.Price > 0]
```

In [9]:
```
len(futbin_data[futbin_data.Price == 0])
```

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

2. Clean data

We need only numeric data, so first of all I'll need to investigate the columns, extract the numeric fields, like the height or the body weight.

Then create dummy columns for the categorical columns like the run style.

At the last step, I'll create some derived columns where I need to merge some information from columns with a lot of different text values. Like the club or the card version. These columns have a lot of different, distinct values, so I will select some of the most valueable categories and will create a special coulmn with 1/0 values. 1 for the most special categories and 0 for the other columns.

Price_Variation

In [10]:
```
futbin_data.Price_Variation.value_counts()
```

Out[10]:

0	3613
2	844
1	596
12.50%	154
3	139
5.88%	92
10.00%	42
11.11%	39
18.18%	17
16.67%	16
4	16
6.25%	15
4.76%	13
9.09%	13
5.26%	13
5.56%	12
2.33%	11
7.14%	11
8.33%	10
6.67%	9
5.00%	9
2.44%	9
14.29%	8
2.22%	8
9.52%	8
2.08%	7
28.57%	7
35.71%	7
3.33%	7
2.13%	7
2.53%	7
3.13%	7
7.69%	7
3.70%	7
25.00%	7
3.51%	7
0.75%	7
6.10%	7
20.00%	7
22.22%	6
4.17%	6
4.08%	6
0.71%	6
11.76%	6
2.67%	5
5.08%	5
2.63%	5
0.74%	5
1.35%	5
4.00%	5
1.67%	5
8.24%	5
1.49%	5
26.67%	5
3.16%	5
2.50%	5
0.73%	5
1.25%	5
33.33%	5
3.85%	5

```
7.50%          5
10.81%         4
15.79%         4
1.69%          4
21.43%         4
4.55%          4
14.81%         4
3.66%          4
18.75%         4
10.64%         4
1.10%          4
8.57%          4
6.52%          4
40.00%         4
2.38%          4
43.06%         4
13.33%         4
3.88%          4
1.89%          4
0.76%          4
6.98%          4
3.11%          4
4.88%          3
15.00%         3
3.03%          3
13.82%         3
6.82%          3
2.27%          3
2.17%          3
2.56%          3
3.57%          3
1.96%          3
2.97%          3
3.80%          3
2.07%          3
3.17%          3
3.23%          3
13.79%         3
2.25%          3
13.04%         3
4.48%          3
15.38%         3
4.05%          3
3.95%          3
32.14%         3
257.14%        3
1.92%          3
3.06%          3
6.41%          3
10.34%         3
11.96%         2
8.45%          2
10.45%         2
2.83%          2
5.04%          2
7.45%          2
2.96%          2
2.23%          2
12.16%         2
1.52%          2
```

4.41%	2
1.06%	2
5.17%	2
4.24%	2
8.75%	2
7.32%	2
7.41%	2
0.78%	2
8.18%	2
1.85%	2
7.58%	2
1.19%	2
3.50%	2
2.86%	2
5.66%	2
10.91%	2
3.94%	2
4.65%	2
17.65%	2
0.56%	2
30.43%	2
0.72%	2
9.80%	2
36.84%	2
9.30%	2
4.69%	2
1.88%	2
1.32%	2
27.78%	2
3.27%	2
0.80%	2
5.68%	2
29.41%	2
0.90%	2
2.70%	2
2.12%	2
1.50%	2
37.04%	2
0.93%	2
8.00%	2
0.69%	2
39.29%	2
11.90%	2
4.35%	2
15.56%	2
6.04%	2
9.62%	2
41.18%	2
4.26%	2
1.77%	2
5.45%	2
5.41%	2
15.15%	2
37.50%	1
42.11%	1
1.22%	1
3.53%	1
10.53%	1
1.59%	1
13.61%	1

19.15%	1
18.00%	1
30.77%	1
1.33%	1
2.04%	1
1.05%	1
12.28%	1
43.33%	1
1,900.00%	1
0.82%	1
23.26%	1
4.29%	1
31.82%	1
13.16%	1
6.90%	1
8.62%	1
35.29%	1
19.77%	1
18.60%	1
14.63%	1
40.91%	1
31.25%	1
23.53%	1
14.52%	1
1.94%	1
8.89%	1
15.22%	1
21.13%	1
7.02%	1
8.77%	1
12.63%	1
4.44%	1
13.21%	1
22.86%	1
10.29%	1
1.75%	1
26.92%	1
9.76%	1
55.88%	1
3.82%	1
8.08%	1
22.73%	1
29.79%	1
30.23%	1
5.13%	1
42.86%	1
26.32%	1
6.80%	1
5.33%	1
6.38%	1
4.62%	1
0.96%	1
3.49%	1
5.92%	1
0.65%	1
7.28%	1
4.46%	1
3.60%	1
9.73%	1
0.32%	1

2.58%	1
14.84%	1
2.26%	1
5.48%	1
0.58%	1
6.20%	1
3.09%	1
4.59%	1
3.35%	1
14.89%	1
0.64%	1
0.67%	1
1.43%	1
4.86%	1
1.87%	1
4.22%	1
1.18%	1
2.65%	1
3.25%	1
0.44%	1
10.58%	1
4.20%	1
0.63%	1
4.72%	1
10.43%	1
0.49%	1
0.30%	1
0.54%	1
5.15%	1
1.74%	1
2.92%	1
1.51%	1
4.57%	1
6.96%	1
5.07%	1
1.64%	1
3.36%	1
1.20%	1
7.16%	1
3.59%	1
0.23%	1
0.66%	1
1.38%	1
0.97%	1
1.83%	1
7.04%	1
2.57%	1
2.54%	1
1.17%	1
7.60%	1
2.11%	1
1.48%	1
7.98%	1
1.71%	1
4.38%	1
13.29%	1
3.68%	1
1.00%	1
0.70%	1
1.14%	1

```
1.15%           1
1.13%           1
0.45%           1
1.40%           1
1.54%           1
5.22%           1
6.42%           1
7.35%           1
8.82%           1
9.38%           1
11.07%          1
10.71%          1
1.03%           1
6.60%           1
7.93%           1
5.96%           1
2.94%           1
2.41%           1
5.60%           1
7.94%           1
1.41%           1
9.86%           1
7.89%           1
8.51%           1
0.37%           1
0.86%           1
4.67%           1
1.11%           1
5.05%           1
3.97%           1
2.84%           1
0.52%           1
2.20%           1
5.24%           1
1.60%           1
16.44%          1
0.55%           1
12.86%          1
1.46%           1
3.90%           1
4.03%           1
2.68%           1
0.68%           1
8.21%           1
14.16%          1
6.29%           1
4.73%           1
9.40%           1
1.79%           1
1.63%           1
1.12%           1
1.08%           1
3.61%           1
3.83%           1
4.82%           1
17.53%          1
0.81%           1
614.29%         1
Name: Price_Variation, dtype: int64
```

too much missing valuees, and based on the differences, it seems to me that this column has a lot of outliners as well, I think it is better to remove it from the dataset.

In [11]:
```
futbin_data = futbin_data.drop(['Price_Variation'], 1)
```

Height

In [12]:
```
futbin_data.head()["Height"][0]
```

Out[12]:
```
'173cm | 5\'8"'
```

In [13]:
```
futbin_data.Height.info()
```

```
<class 'pandas.core.series.Series'>
Int64Index: 6409 entries, 0 to 7159
Series name: Height
Non-Null Count  Dtype
--------------  -----
6409 non-null   object
dtypes: object(1)
memory usage: 100.1+ KB
```

In [14]:
```
futbin_data.Height.value_counts()
```

Out[14]:
```
180cm | 5'11"    481
185cm | 6'1"     473
178cm | 5'10"    419
183cm | 6'0"     375
175cm | 5'9"     370
188cm | 6'2"     330
182cm | 6'0"     293
186cm | 6'1"     277
184cm | 6'0"     274
181cm | 5'11"    244
187cm | 6'2"     243
176cm | 5'9"     230
177cm | 5'10"    227
179cm | 5'10"    221
190cm | 6'3"     220
173cm | 5'8"     204
191cm | 6'3"     176
174cm | 5'9"     154
189cm | 6'2"     152
170cm | 5'7"     145
172cm | 5'8"     135
192cm | 6'4"     128
193cm | 6'4"     107
171cm | 5'7"      84
194cm | 6'4"      77
168cm | 5'6"      66
195cm | 6'5"      51
169cm | 5'7"      47
196cm | 6'5"      43
167cm | 5'6"      30
166cm | 5'5"      23
197cm | 6'6"      22
198cm | 6'6"      22
165cm | 5'5"      19
163cm | 5'4"       9
201cm | 6'7"       7
164cm | 5'5"       7
200cm | 6'7"       5
199cm | 6'6"       5
162cm | 5'4"       4
206cm | 6'9"       3
204cm | 6'8"       2
161cm | 5'3"       2
202cm | 6'8"       1
158cm | 5'2"       1
160cm | 5'3"       1
Name: Height, dtype: int64
```

In [15]:
```
# remove 0 values, 4 records
```

In [16]:
```
futbin_data.shape[0]
```

Out[16]:
```
6409
```

In [17]:
```
futbin_data[futbin_data.Height == "0"]
```

Out[17]:

	Name	Club	Nation	League	Rating	Main_Position	Alternate_Positions	Card_Version	Run_Style

In [18]:
```
futbin_data = futbin_data.drop(futbin_data[futbin_data.Height == "0"].index)
```

In [19]:
```
futbin_data.shape[0]
```

Out[19]:
```
6409
```

In [20]:
```
futbin_data[futbin_data.Height == "0"]
```

Out[20]:

	Name	Club	Nation	League	Rating	Main_Position	Alternate_Positions	Card_Version	Run_Style

In [21]:
```
futbin_data['Height_in_cm'] = futbin_data['Height'].str[:3]
```

In [22]:
```
futbin_data['Height_in_cm'].info()
```

```
<class 'pandas.core.series.Series'>
Int64Index: 6409 entries, 0 to 7159
Series name: Height_in_cm
Non-Null Count  Dtype
--------------  -----
6409 non-null   object
dtypes: object(1)
memory usage: 100.1+ KB
```

In [23]:
```
futbin_data['Height_in_cm'] = pd.to_numeric(futbin_data['Height_in_cm'])
```

In [24]:
```
futbin_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 6409 entries, 0 to 7159
Data columns (total 26 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Name                    6409 non-null   object 
 1   Club                    6409 non-null   object 
 2   Nation                  6409 non-null   object 
 3   League                  6409 non-null   object 
 4   Rating                  6409 non-null   int64  
 5   Main_Position           6409 non-null   object 
 6   Alternate_Positions     6409 non-null   object 
 7   Card_Version            6409 non-null   object 
 8   Run_Style               6409 non-null   object 
 9   Price                   6409 non-null   float64
 10  Skills_Star             6409 non-null   int64  
 11  Weak_Foot_Star          6409 non-null   int64  
 12  Attack_Workrate         6409 non-null   object 
 13  Defense_Workrate        6409 non-null   object 
 14  Pace / Diving           6409 non-null   float64
 15  Shooting / Handling     6409 non-null   int64  
 16  Passing / Kicking       6409 non-null   int64  
 17  Dribbling / Reflexes    6409 non-null   int64  
 18  Defense / Speed         6409 non-null   int64  
 19  Physical / Positioning  6409 non-null   int64  
 20  Height                  6409 non-null   object 
 21  BodyType                6409 non-null   object 
 22  Popularity              6409 non-null   int64  
 23  Base_Stats              6409 non-null   int64  
 24  Ingame_Stats            6409 non-null   int64  
 25  Height_in_cm            6409 non-null   int64  
dtypes: float64(2), int64(12), object(12)
memory usage: 1.3+ MB
```

Body type

In [25]:
```
futbin_data.BodyType.info()
```

```
<class 'pandas.core.series.Series'>
Int64Index: 6409 entries, 0 to 7159
Series name: BodyType
Non-Null Count  Dtype  
--------------  -----  
6409 non-null   object 
dtypes: object(1)
memory usage: 100.1+ KB
```

In [26]:
```
futbin_data.BodyType.value_counts()
```

Out[26]:

Average (70kg)	223
Average (75kg)	219
Average (73kg)	178
Average (72kg)	174
High & Average (80kg)	171
Average (74kg)	160
Lean (70kg)	148
Average (76kg)	137
Average (77kg)	135
Lean (72kg)	123
High & Average (78kg)	121
High & Average (82kg)	119
Average (71kg)	118
Lean (73kg)	113
High & Average (83kg)	112
High & Average (85kg)	109
Lean (68kg)	106
Average (78kg)	106
Lean (75kg)	102
Average (68kg)	99
Average (79kg)	97
High & Average (84kg)	96
Lean (74kg)	95
Average (80kg)	90
High & Average (81kg)	87
Lean (67kg)	82
Lean (65kg)	81
High & Average (79kg)	75
High & Average+ (80kg)	73
Average (69kg)	71
Lean (71kg)	69
High & Average+ (76kg)	66
High & Average (75kg)	66
High & Average (77kg)	64
Lean (66kg)	64
High & Average+ (75kg)	63
High & Average (88kg)	63
High & Average (86kg)	59
High & Average (76kg)	58
Lean (69kg)	56
High & Average+ (78kg)	55
Average (67kg)	53
High & Average+ (77kg)	48
High & Average (87kg)	46
Lean (64kg)	42
Average (65kg)	42
High & Average+ (82kg)	41
Lean (76kg)	41
High & Average+ (79kg)	38
High & Average+ (74kg)	38
High & Average (90kg)	36
Average (81kg)	35
Lean (77kg)	35
High & Average+ (73kg)	34
Average (82kg)	33
Average (83kg)	31
High & Average (89kg)	31
High & Average (74kg)	29
Lean (78kg)	29
Lean (63kg)	27

```
Average  (66kg)                26
High & Average+  (85kg)        26
High & Average+  (83kg)        25
Lean  (62kg)                   25
High & Average+  (81kg)        25
High & Average+  (70kg)        25
Short & Lean-  (65kg)          24
Average  (64kg)                21
Lean  (79kg)                   21
Short & Lean  (65kg)           21
Average  (84kg)                20
Stocky  (75kg)                 20
High & Average  (91kg)         20
Short & Lean-  (64kg)          19
High & Average+  (84kg)        19
High & Average+  (72kg)        19
High & Average  (72kg)         17
High & Average  (73kg)         17
Stocky  (79kg)                 17
Average  (63kg)                16
High & Average+  (71kg)        16
Short & Lean  (63kg)           15
Lean  (60kg)                   15
Short & Lean-  (69kg)          15
High & Average  (92kg)         14
High & Stocky  (80kg)          14
High & Stocky  (90kg)          14
Average  (85kg)                14
Short & Lean  (62kg)           14
Short & Lean  (60kg)           14
Stocky  (74kg)                 14
Short & Lean-  (66kg)          14
Stocky  (80kg)                 13
Short & Lean-  (68kg)          13
High & Average  (93kg)         13
Short & Lean-  (67kg)          13
Short & Lean-  (70kg)          13
Unique  (70kg)                 13
Stocky  (76kg)                 12
Unique  (73kg)                 12
Normal                         12
High & Average+  (68kg)        11
Stocky  (82kg)                 11
Unique  (78kg)                 11
High & Average  (70kg)         11
Short & Lean-  (61kg)          11
Short & Lean-  (60kg)          11
Lean  (80kg)                   11
Stocky  (72kg)                 10
Short & Lean-  (63kg)          10
Stocky  (78kg)                 10
High & Average+  (86kg)        10
Short and Balanced   (68kg)    10
Average  (62kg)                10
High & Stocky  (85kg)          10
High & Average  (94kg)         10
High & Stocky  (86kg)          10
Unique  (75kg)                  9
Stocky  (77kg)                  9
Short & Lean  (61kg)            8
```

```
Unique  (69kg)                  8
Unique  (76kg)                  8
High & Stocky  (88kg)           8
Short & Lean  (58kg)            8
Stocky  (73kg)                  8
High & Average+  (90kg)         8
Stocky  (85kg)                  8
Average  (60kg)                 7
High & Average  (97kg)          7
Short & Lean-  (71kg)           7
Lean  (61kg)                    7
Short & Lean  (64kg)            7
Stocky  (83kg)                  7
High & Average+  (87kg)         7
Stocky  (70kg)                  7
Unique  (80kg)                  6
Messi  (67kg)                   6
Stocky  (69kg)                  6
Short & Lean  (66kg)            6
Stocky  (84kg)                  6
High & Stocky  (89kg)           6
Stocky  (87kg)                  6
Short & Lean  (57kg)            6
Short and Balanced   (65kg)     6
Short & Lean-  (62kg)           6
High & Average  (95kg)          5
High & Average+  (88kg)         5
Lean  (58kg)                    5
High & Stocky  (87kg)           5
High & Average  (98kg)          5
Unique  (74kg)                  5
Stocky  (71kg)                  5
High & Average+  (69kg)         5
High & Stocky  (78kg)           5
Average  (87kg)                 5
Short and Balanced   (70kg)     5
Short & Lean  (67kg)            5
High & Stocky  (83kg)           5
Unique  (67kg)                  5
Unique  (89kg)                  5
High & Average+  (93kg)         5
High & Stocky  (84kg)           4
Unique  (94kg)                  4
Short & Lean  (68kg)            4
High & Stocky  (81kg)           4
High & Stocky  (79kg)           4
Short & Lean  (69kg)            4
Stocky  (81kg)                  4
High & Average  (68kg)          4
High & Average+  (89kg)         4
High & Average+  (67kg)         4
High & Average+  (92kg)         4
High & Average  (71kg)          4
Unique  (71kg)                  4
Unique  (81kg)                  4
Short & Lean-  (73kg)           4
Unique  (68kg)                  4
Lean  (59kg)                    4
High & Stocky  (82kg)           3
Unique  (77kg)                  3
```

```
Unique  (84kg)                  3
High & Stocky  (92kg)           3
High & Stocky  (96kg)           3
Lean  (85kg)                    3
Unique  (82kg)                  3
Short & Lean  (72kg)            3
Short & Lean  (59kg)            3
Stocky  (92kg)                  3
Stocky  (86kg)                  3
Short and Balanced   (72kg)     3
Short & Lean  (70kg)            3
Unique  (85kg)                  3
High & Average+  (91kg)         3
Short and Balanced   (71kg)     3
High & Stocky  (98kg)           3
Short & Lean-  (75kg)           2
Average  (88kg)                 2
High & Stocky  (76kg)           2
Stocky  (68kg)                  2
Lean  (86kg)                    2
Short & Lean  (56kg)            2
Short & Lean-  (74kg)           2
High & Stocky  (100kg)          2
High & Average+  (64kg)         2
Lean  (82kg)                    2
Short & Lean  (71kg)            2
High & Stocky  (75kg)           2
Short & Lean-  (55kg)           2
Unique  (90kg)                  2
Short & Lean-  (76kg)           2
Lean  (56kg)                    2
Short & Lean-  (72kg)           2
Stocky  (90kg)                  2
Unique  (66kg)                  2
Stocky  (67kg)                  2
High & Stocky  (94kg)           2
Short & Lean-  (58kg)           2
Unique  (60kg)                  2
Salah  (71kg)                   2
R9  (78kg)                      2
Short & Lean-  (59kg)           2
Unique  (64kg)                  2
High & Stocky  (99kg)           2
Unique  (72kg)                  2
High & Average  (69kg)          2
High & Lean   (75kg)            2
Short & Lean-  (77kg)           2
High & Stocky  (91kg)           2
Short and Balanced   (73kg)     2
High & Stocky  (93kg)           2
Unique  (87kg)                  2
Short & Lean  (77kg)            1
Short & Lean  (54kg)            1
High & Average+  (65kg)         1
Short & Lean-  (57kg)           1
High & Average  (99kg)          1
Short & Lean-  (49kg)           1
Lean  (81kg)                    1
Short and Balanced   (66kg)     1
Stocky  (66kg)                  1
```

```
High & Average  (102kg)          1
Average  (93kg)                  1
Average  (57kg)                  1
Short and Balanced   (67kg)      1
Stocky  (94kg)                   1
Short and Balanced   (75kg)      1
High & Stocky  (77kg)            1
Lean  (90kg)                     1
High & Average  (96kg)           1
Short & Lean  (73kg)             1
Unique  (86kg)                   1
Unique  (91kg)                   1
High & Average  (100kg)          1
Ronaldinho  (78kg)               1
High & Lean   (93kg)             1
Short and Balanced   (69kg)      1
Unique  (61kg)                   1
Unique  (59kg)                   1
High & Lean   (80kg)             1
Shaqiri  (72kg)                  1
High & Lean   (77kg)             1
Ronaldinho  (76kg)               1
Average  (0kg)                   1
Stocky  (88kg)                   1
High & Average+  (66kg)          1
High & Average  (67kg)           1
Lean  (84kg)                     1
Average  (89kg)                  1
Unique  (92kg)                   1
Short and Balanced   (62kg)      1
High & Stocky  (74kg)            1
Stocky  (64kg)                   1
CR7  (83kg)                      1
Courtois  (96kg)                 1
Unique  (93kg)                   1
Stocky  (65kg)                   1
High & Average  (65kg)           1
Average  (86kg)                  1
Short and Balanced   (63kg)      1
Average  (91kg)                  1
High & Stocky  (95kg)            1
Stocky  (89kg)                   1
Name: BodyType, dtype: int64
```

I need to create two more columns from this one. The first one would be a numeric field in kg. The second is a categorical field with the body type.

In [27]:
```
futbin_data[['BodyType_Text', 'BodyType_Weight']] = futbin_data['BodyType'].str.split(
```

First try, using ' ' (double spaces), let's try, is it working?

In [28]:
```
futbin_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 6409 entries, 0 to 7159
Data columns (total 28 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Name                    6409 non-null   object 
 1   Club                    6409 non-null   object 
 2   Nation                  6409 non-null   object 
 3   League                  6409 non-null   object 
 4   Rating                  6409 non-null   int64  
 5   Main_Position           6409 non-null   object 
 6   Alternate_Positions     6409 non-null   object 
 7   Card_Version            6409 non-null   object 
 8   Run_Style               6409 non-null   object 
 9   Price                   6409 non-null   float64
 10  Skills_Star             6409 non-null   int64  
 11  Weak_Foot_Star          6409 non-null   int64  
 12  Attack_Workrate         6409 non-null   object 
 13  Defense_Workrate        6409 non-null   object 
 14  Pace / Diving           6409 non-null   float64
 15  Shooting / Handling     6409 non-null   int64  
 16  Passing / Kicking       6409 non-null   int64  
 17  Dribbling / Reflexes    6409 non-null   int64  
 18  Defense / Speed         6409 non-null   int64  
 19  Physical / Positioning  6409 non-null   int64  
 20  Height                  6409 non-null   object 
 21  BodyType                6409 non-null   object 
 22  Popularity              6409 non-null   int64  
 23  Base_Stats              6409 non-null   int64  
 24  Ingame_Stats            6409 non-null   int64  
 25  Height_in_cm            6409 non-null   int64  
 26  BodyType_Text           6409 non-null   object 
 27  BodyType_Weight         6397 non-null   object 
dtypes: float64(2), int64(12), object(14)
memory usage: 1.4+ MB
```

Some rows are missing from the Weight coulmn, maybe double-space is not the right separator for every row.

In [29]:
```
futbin_data.BodyType_Text.value_counts()
```

Out[29]:
```
Average               2128
High & Average        1477
Lean                  1313
High & Average+        681
Stocky                 191
Short & Lean-          178
Unique                 128
Short & Lean           128
High & Stocky          118
Short and Balanced      35
Normal                  12
Messi                    6
High & Lean              5
Salah                    2
Ronaldinho               2
R9                       2
CR7                      1
Courtois                 1
Shaqiri                  1
Name: BodyType_Text, dtype: int64
```

In [30]:
```
futbin_data.BodyType_Weight.value_counts()
```

Out[30]:

(75kg)	483
(70kg)	443
(80kg)	378
(73kg)	367
(72kg)	351
(74kg)	344
(78kg)	340
(76kg)	327
(77kg)	298
(79kg)	252
(68kg)	243
(71kg)	227
(82kg)	212
(83kg)	181
(85kg)	173
(67kg)	171
(65kg)	171
(69kg)	167
(81kg)	160
(84kg)	149
(66kg)	114
(64kg)	94
(86kg)	86
(88kg)	79
(87kg)	71
(63kg)	68
(90kg)	63
(62kg)	55
(60kg)	49
(89kg)	48
(91kg)	27
(61kg)	27
(92kg)	25
(93kg)	22
(94kg)	17
(58kg)	15
(59kg)	10
(68kg)	10
(98kg)	8
(57kg)	8
(97kg)	7
(65kg)	6
(95kg)	6
(96kg)	5
(70kg)	5
(56kg)	4
(100kg)	3
(71kg)	3
(99kg)	3
(75kg)	3
(72kg)	3
(73kg)	2
(55kg)	2
(66kg)	1
(102kg)	1
(49kg)	1
(67kg)	1
(63kg)	1
(80kg)	1
(62kg)	1

```
(0kg)        1
 (77kg)      1
 (69kg)      1
 (93kg)      1
(54kg)       1
Name: BodyType_Weight, dtype: int64
```

Second try, sometimes we have 2 or 3 spaces as a separator so I will try a parenthesis (

In [31]:
```
futbin_data[['BodyType_Text', 'BodyType_Weight']] = futbin_data['BodyType'].str.split(
```

In [32]:
```
futbin_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 6409 entries, 0 to 7159
Data columns (total 28 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   Name                   6409 non-null   object 
 1   Club                   6409 non-null   object 
 2   Nation                 6409 non-null   object 
 3   League                 6409 non-null   object 
 4   Rating                 6409 non-null   int64  
 5   Main_Position          6409 non-null   object 
 6   Alternate_Positions    6409 non-null   object 
 7   Card_Version           6409 non-null   object 
 8   Run_Style              6409 non-null   object 
 9   Price                  6409 non-null   float64
 10  Skills_Star            6409 non-null   int64  
 11  Weak_Foot_Star         6409 non-null   int64  
 12  Attack_Workrate        6409 non-null   object 
 13  Defense_Workrate       6409 non-null   object 
 14  Pace / Diving          6409 non-null   float64
 15  Shooting / Handling    6409 non-null   int64  
 16  Passing / Kicking      6409 non-null   int64  
 17  Dribbling / Reflexes   6409 non-null   int64  
 18  Defense / Speed        6409 non-null   int64  
 19  Physical / Positioning 6409 non-null   int64  
 20  Height                 6409 non-null   object 
 21  BodyType               6409 non-null   object 
 22  Popularity             6409 non-null   int64  
 23  Base_Stats             6409 non-null   int64  
 24  Ingame_Stats           6409 non-null   int64  
 25  Height_in_cm           6409 non-null   int64  
 26  BodyType_Text          6409 non-null   object 
 27  BodyType_Weight        6397 non-null   object 
dtypes: float64(2), int64(12), object(14)
memory usage: 1.4+ MB
```

In [33]:
```
futbin_data.BodyType_Text.value_counts()
```

Out[33]:
```
Average                 2128
High & Average          1477
Lean                    1313
High & Average+          681
Stocky                   191
Short & Lean-            178
Unique                   128
Short & Lean             128
High & Stocky            118
Short and Balanced        35
Normal                    12
Messi                      6
High & Lean                5
Salah                      2
Ronaldinho                 2
R9                         2
CR7                        1
Courtois                   1
Shaqiri                    1
Name: BodyType_Text, dtype: int64
```

In [34]:
```
futbin_data.BodyType_Weight.value_counts()
```

Out[34]:
```
75kg)     486
70kg)     448
80kg)     379
73kg)     369
72kg)     354
74kg)     344
78kg)     340
76kg)     327
77kg)     299
68kg)     253
79kg)     252
71kg)     230
82kg)     212
83kg)     181
65kg)     177
85kg)     173
67kg)     172
69kg)     168
81kg)     160
84kg)     149
66kg)     115
64kg)      94
86kg)      86
88kg)      79
87kg)      71
63kg)      69
90kg)      63
62kg)      56
60kg)      49
89kg)      48
61kg)      27
91kg)      27
92kg)      25
93kg)      23
94kg)      17
58kg)      15
59kg)      10
57kg)       8
98kg)       8
97kg)       7
95kg)       6
96kg)       5
56kg)       4
100kg)      3
99kg)       3
55kg)       2
0kg)        1
49kg)       1
102kg)      1
54kg)       1
Name: BodyType_Weight, dtype: int64
```

In [35]:
```
futbin_data.BodyType_Weight.count()
```

Out[35]:
```
6397
```

In [36]:
```
futbin_data[futbin_data.BodyType_Weight.isnull() == True]
```

Out[36]:

	Name	Club	Nation	League	Rating	Main_Position	Alternate_Positions	Card_V
21	Müller	FUT ICONS	Germany	Icons	94	ST	CF	
45	Jairzinho	FUT ICONS	Brazil	Icons	93	RW	0	
65	Jairzinho	FUT ICONS	Brazil	Icons	92	RW	RM,ST	
243	Ledley King	HERO	England	Premier League	89	CB	0	Fantas
516	Ledley King	HERO	England	Premier League	86	CB	0	
836	Lucas Ocampos	Ajax	Argentina	Eredivisie	81	LW	LM,RW	I
992	Martin Dúbravka	Newcastle Utd	Slovakia	Premier League	80	GK	0	nc
1586	Jan Bednarek	Southampton	Poland	Premier League	76	CB	0	nc
4769	Sylvester Jasper	Fulham	Bulgaria	Premier League	65	RM	LM,RW	nc
5506	Harvey Vale	Chelsea	England	Premier League	63	LW	LM,CAM	nc
6471	Mamadou Camara	RC Lens	Senegal	Ligue 1	59	CM	0	nc
6474	Chinonso Offor	Impact Montréal	Nigeria	Major League Soccer	59	ST	LM,CF	nc

In [37]:
```
futbin_data = futbin_data.drop(futbin_data[futbin_data.BodyType_Weight.isnull() == Tru
```

In [38]:
```
futbin_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 6397 entries, 0 to 7159
Data columns (total 28 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Name                    6397 non-null   object
 1   Club                    6397 non-null   object
 2   Nation                  6397 non-null   object
 3   League                  6397 non-null   object
 4   Rating                  6397 non-null   int64
 5   Main_Position           6397 non-null   object
 6   Alternate_Positions     6397 non-null   object
 7   Card_Version            6397 non-null   object
 8   Run_Style               6397 non-null   object
 9   Price                   6397 non-null   float64
 10  Skills_Star             6397 non-null   int64
 11  Weak_Foot_Star          6397 non-null   int64
 12  Attack_Workrate         6397 non-null   object
 13  Defense_Workrate        6397 non-null   object
 14  Pace / Diving           6397 non-null   float64
 15  Shooting / Handling     6397 non-null   int64
 16  Passing / Kicking       6397 non-null   int64
 17  Dribbling / Reflexes    6397 non-null   int64
 18  Defense / Speed         6397 non-null   int64
 19  Physical / Positioning  6397 non-null   int64
 20  Height                  6397 non-null   object
 21  BodyType                6397 non-null   object
 22  Popularity              6397 non-null   int64
 23  Base_Stats              6397 non-null   int64
 24  Ingame_Stats            6397 non-null   int64
 25  Height_in_cm            6397 non-null   int64
 26  BodyType_Text           6397 non-null   object
 27  BodyType_Weight         6397 non-null   object
dtypes: float64(2), int64(12), object(14)
memory usage: 1.4+ MB
```

In [39]:
```
futbin_data.BodyType_Text.value_counts()
```

Out[39]:
```
Average                2128
High & Average         1477
Lean                   1313
High & Average+         681
Stocky                  191
Short & Lean-           178
Short & Lean            128
Unique                  128
High & Stocky           118
Short and Balanced       35
Messi                     6
High & Lean               5
Ronaldinho                2
Salah                     2
R9                        2
CR7                       1
Courtois                  1
Shaqiri                   1
Name: BodyType_Text, dtype: int64
```

In [40]:
```
futbin_data.BodyType_Text = futbin_data.BodyType_Text.str.strip()
```

In [41]:
```
futbin_data.BodyType_Weight.value_counts()
```

Out[41]:
```
75kg)     486
70kg)     448
80kg)     379
73kg)     369
72kg)     354
74kg)     344
78kg)     340
76kg)     327
77kg)     299
68kg)     253
79kg)     252
71kg)     230
82kg)     212
83kg)     181
65kg)     177
85kg)     173
67kg)     172
69kg)     168
81kg)     160
84kg)     149
66kg)     115
64kg)      94
86kg)      86
88kg)      79
87kg)      71
63kg)      69
90kg)      63
62kg)      56
60kg)      49
89kg)      48
61kg)      27
91kg)      27
92kg)      25
93kg)      23
94kg)      17
58kg)      15
59kg)      10
57kg)       8
98kg)       8
97kg)       7
95kg)       6
96kg)       5
56kg)       4
100kg)      3
99kg)       3
55kg)       2
0kg)        1
49kg)       1
102kg)      1
54kg)       1
Name: BodyType_Weight, dtype: int64
```

In [42]:
```
futbin_data.BodyType_Weight = futbin_data.BodyType_Weight.str[:-3]
```

In [43]:
```
futbin_data.BodyType_Weight = pd.to_numeric(futbin_data.BodyType_Weight)
```

In [44]:
```
futbin_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 6397 entries, 0 to 7159
Data columns (total 28 columns):
 #   Column                 Non-Null Count  Dtype
---  ------                 --------------  -----
 0   Name                   6397 non-null   object
 1   Club                   6397 non-null   object
 2   Nation                 6397 non-null   object
 3   League                 6397 non-null   object
 4   Rating                 6397 non-null   int64
 5   Main_Position          6397 non-null   object
 6   Alternate_Positions    6397 non-null   object
 7   Card_Version           6397 non-null   object
 8   Run_Style              6397 non-null   object
 9   Price                  6397 non-null   float64
 10  Skills_Star            6397 non-null   int64
 11  Weak_Foot_Star         6397 non-null   int64
 12  Attack_Workrate        6397 non-null   object
 13  Defense_Workrate       6397 non-null   object
 14  Pace / Diving          6397 non-null   float64
 15  Shooting / Handling    6397 non-null   int64
 16  Passing / Kicking      6397 non-null   int64
 17  Dribbling / Reflexes   6397 non-null   int64
 18  Defense / Speed        6397 non-null   int64
 19  Physical / Positioning 6397 non-null   int64
 20  Height                 6397 non-null   object
 21  BodyType               6397 non-null   object
 22  Popularity             6397 non-null   int64
 23  Base_Stats             6397 non-null   int64
 24  Ingame_Stats           6397 non-null   int64
 25  Height_in_cm           6397 non-null   int64
 26  BodyType_Text          6397 non-null   object
 27  BodyType_Weight        6397 non-null   int64
dtypes: float64(2), int64(13), object(13)
memory usage: 1.4+ MB
```

In [45]:
```
futbin_data.head()
```

Out[45]:

	Name	Club	Nation	League	Rating	Main_Position	Alternate_Positions	Card_Version	R
0	Pelé	FUT ICONS	Brazil	Icons	98	CAM	CF,ST	Icon	
1	Lionel Messi	Paris SG	Argentina	Ligue 1	98	RW	RM	TOTY	C
2	Lionel Messi	Paris SG	Argentina	Ligue 1	98	ST	RM,RW	TOTS	C
3	Karim Benzema	Real Madrid	France	LaLiga Santander	97	CF	ST	TOTY	C
4	Kylian Mbappé	Paris SG	France	Ligue 1	97	ST	CF,LW	TOTY	C

Remove extracted columns

In [46]:
```
futbin_data = futbin_data.drop(['Height','BodyType'], 1)
```

In [47]:
```
futbin_data.Club.value_counts()
```

Out[47]:

FUT ICONS	128
HERO	38
Arsenal	29
Liverpool	25
Paris SG	25
Manchester City	24
Manchester Utd	24
Real Madrid	23
Dortmund	22
RC Celta	22
Banfield	21
FC Bayern	21
Atlético de Madrid	21
Juventus	20
Austin FC	20
Vélez Sarsfield	20
FC Barcelona	20
Ajax	19
Roma FC	19
Unión	19
Newcastle Utd	18
RB Leipzig	18
Napoli	18
Zagłębie Lubin	18
Suwon Samsung	18
Sevilla FC	18
AS Monaco	18
Chelsea	18
OM	18
LAFC	17
Sporting CP	17
FC Augsburg	17
Milan	17
Racing Club	17
Inter	17
Lanús	17
Athletic Club	17
Cádiz CF	17
Real Sociedad	16
LOSC	16
FC Schalke 04	16
RC Lens	16
Cheltenham Town	16
Strasbourg	16
Getafe CF	16
Leverkusen	16
Karlsruher SC	16
La Spezia	15
Burnley	15
Torino	15
Stade Brestois 29	15
Independiente	15
1. FC Heidenheim	15
Granada CF	15
RCD Mallorca	15
Libertad	15
Troyes	15
FC Seoul	15
SS Lazio	15
Swansea City	15

Wycombe	15
Rennes	15
FCSB	15
Spurs	15
Hertha Berlin	15
Frankfurt	14
VfB Stuttgart	14
Huddersfield	14
Al Shabab	14
Dep. Táchira	14
Hull City	14
Aston Villa	14
Beşiktaş	14
Malmö FF	14
Elche CF	14
Talleres	14
SL Benfica	14
Varbergs BoIS	14
Jeonbuk Hyundai	14
Valencia CF	14
Girona FC	14
Wolves	14
Western United	14
Greuther Fürth	14
QPR	14
Boavista FC	14
Guangzhou	14
Cambridge Utd	14
AC Monza	14
Bologna	14
SC Freiburg	14
Union Berlin	14
Melb. Victory	14
Toronto FC	13
CS Emelec	13
SC Cambuur	13
Al Ain FC	13
SK Sturm Graz	13
Leicester City	13
OGC Nice	13
Feyenoord	13
AFC Wimbledon	13
Cerro Porteño	13
Blackburn Rovers	13
Dynamo Dresden	13
Antalyaspor	13
FC Rapid	13
VfL Bochum	13
D.C. United	13
Reading	13
FC København	13
PSV	13
Coventry City	13
FC Nordsjælland	13
Deportes Antofagasta	13
Damac FC	13
APOEL Nicosia	13
FC Nantes	13
Hellas Verona	13
FC Basel 1893	13

Sampdoria	13
Stoke City	13
CA Osasuna	13
Al Taawoun	13
St. Pats	13
Wigan Athletic	13
Real Betis	12
Kasimpaşa	12
Wrexham AFC	12
LASK	12
Cracovia	12
Viking FK	12
Club Brugge	12
Grimsby Town	12
Shenzhen Kaisa	12
Molde FK	12
AZ	12
Başakşehir	12
Venezia	12
Kristiansund	12
Colorado Rapids	12
Tranmere Rovers	12
Al-Wehda FC	12
Bristol Rovers	12
Barnsley	12
New York City FC	12
BK Häcken	12
Atalanta BC	12
Leeds United	12
AFC Bournemouth	12
Seattle Sounders	12
Galatasaray	12
Zulte Waregem	12
Al Raed	12
1860 München	12
Meizhou Hakka	12
Rot-Weiss Essen	12
West Brom	12
Stockport	12
Shijiazhuang	12
Oxford United	12
Panathinaikos	12
Forest Green	12
SC Paderborn 07	12
Guangzhou R&F	12
Everton	12
KV Kortrijk	12
Sheffield Utd	12
Al Hilal	12
Daegu FC	12
TSG Hoffenheim	12
Sunderland	12
UD Ibiza	11
Zhejiang Pro	11
Warta Poznań	11
FC Metz	11
CS Mioveni	11
Rodez AF	11
Grenoble	11
KAS Eupen	11

Salford City	11
UTA Arad	11
Sassuolo	11
SV Waldhof	11
Suwon FC	11
Albacete Balompié	11
Górnik Zabrze	11
Spal	11
Dijon FCO	11
Incheon United	11
Seongnam FC	11
Ternana	11
Aarhus AGF	11
Hansa Rostock	11
FC Groningen	11
Pogoń Szczecin	11
Palermo	11
FC Botoşani	11
Śląsk Wrocław	11
IFK Norrköping	11
VfL Wolfsburg	11
St. Johnstone	11
FC Twente	11
Al Nassr	11
Raków Częstochowa	11
MK Dons	11
RKC Waalwijk	11
VfL Osnabrück	11
Como	11
Sepsi OSK	11
Nashville SC	11
Düsseldorf	11
RB Salzburg	11
Accrington	11
1. FC Nürnberg	11
Shamrock Rovers	11
FC Utrecht	11
Univ. Craiova	11
Millwall	11
Lyngby BK	11
Shanghai Shenhua	11
FC Dallas	11
Derry City	11
Philadelphia	11
Norwich	11
SC Braga	11
FC Zürich	11
FC Cincinnati	11
Brighton	11
Changchun Yatai	11
UD Las Palmas	11
Livingston	11
Fulham	11
FC Voluntari	10
Houston Dynamo	10
Colón	10
Crystal Palace	10
Djurgårdens IF	10
New England	10
Kilmarnock	10

Ulsan Hyundai	10
FC Porto	10
M'gladbach	10
Borussia Dortmund II	10
Lecce	10
Slavia Praha	10
FC Cartagena	10
CD Tenerife	10
Cercle Brugge	10
Burton Albion	10
SVWW	10
Legia Warszawa	10
R. Sporting	10
Al-Tai	10
Luton Town	10
AEK Athens	10
Bohemian FC	10
FK Jerv	10
Brescia	10
Newport County	10
Rochdale	10
Grasshoppers	10
Blackpool	10
Peterborough	10
Sparta Praha	10
GD Chaves	10
Celtic	10
IF Elfsborg	10
AJ Auxerre	10
Sutton United	10
SV Elversberg	10
Sligo Rovers	10
Cardiff City	10
Central Córdoba	10
Rotherham Utd	10
Fiorentina	10
FC Petrolul	10
Villarreal CF	10
Fleetwood Town	10
Portland Timbers	10
Chicago Fire	10
Charlotte FC	10
Watford	10
KV Mechelen	10
Shandong Luneng	10
Angers SCO	10
Viborg FF	10
Southampton	10
Mamelodi Sundowns	10
Heart of Midlothian	10
Toulouse FC	10
AC Ajaccio	10
Al Ittihad	10
Bristol City	10
Cremonese	10
FC Argeș	10
Al Adalah	10
FC Univ. Cluj	10
FC St. Gallen	10
Leyton Orient	10

```
SK Austria Klagenfurt               10
R. Oviedo                           10
Empoli                              10
Kalmar FF                            9
Kaiserslautern                       9
Stevenage                            9
Atlanta United                       9
Melbourne City                       9
Club Atlético Sarmiento (Junin)      9
Columbus Crew SC                     9
FC Arouca                            9
Orlando City                         9
Jahn Regensburg                      9
Nott'm Forest                        9
Radomiak Radom                       9
BSC Young Boys                       9
Dalian Pro                           9
Shelbourne                           9
R. Valladolid CF                     9
Gillingham                           9
1. FC Köln                           9
Vålerenga Fotball                    9
Servette FC                          9
Al Khaleej                           9
Modena                               9
Silkeborg IF                         9
Impact Montréal                      9
Piast Gliwice                        9
V. Guimarães                         9
CFR 1907 Cluj                        9
SV Ried                              9
Chindia                              9
OL                                   9
Genoa                                9
R. Zaragoza                          9
SD Eibar                             9
Doncaster                            9
Atlético Tucumán                     9
SC Freiburg II                       9
Boca Juniors                         9
Cittadella                           9
Jeju United                          9
GD Estoril Praia                     9
Estudiantes                          9
Lechia Gdańsk                        9
Wuhan Zall                           9
CD Leganés                           9
Exeter City                          9
FSV Zwickau                          9
Montpellier                          9
OH Leuven                            9
LA Galaxy                            9
KVC Westerlo                         9
MSV Duisburg                         9
Miedź Legnica                        9
Cagliari                             9
Paços Ferreira                       9
Middlesbrough                        9
FCSM                                 9
AC Horsens                           9
```

Shrewsbury	9
Royale Union SG	9
Brentford	9
Royal Antwerp FC	9
Perth Glory	9
Rangers	9
Burgos CF	9
FK Austria Wien	9
Viktoria Plzeň	9
FC Sion	9
FC Midtjylland	9
Sp. Charleroi	9
Al Batin	9
Aalesunds FK	9
West Ham	9
Paris FC	9
FK Bodø/Glimt	9
WSG Tirol	9
Gangwon FC	9
Rosenborg BK	8
Wuhan Three Towns	8
Braunschweig	8
Perugia	8
Harrogate Town	8
SV Meppen	8
Inter Miami CF	8
Saarbrücken	8
Port Vale	8
Südtirol	8
Sandefjord	8
FC Farul Constanta	8
TSV Hartberg	8
Hebei CFFC	8
Crawley Town	8
FC Goa	8
Al Fayha	8
SK Rapid Wien	8
Nîmes Olympique	8
Sarpsborg 08	8
FC Winterthur	8
Hibernian	8
HamKam Fotball	8
Real Salt Lake	8
Walsall	8
Ettifaq FC	8
Strømsgodset IF	8
SV Darmstadt 98	8
Alanyaspor	8
Shanghai SIPG	8
Nacional	8
Aberdeen	8
Hammarby IF	8
Vitesse	8
Niort	8
Bordeaux	8
KAA Gent	8
FC Famalicão	8
Minnesota United	8
RSC Anderlecht	8
Whitecaps FC	8

Godoy Cruz	8
Villarreal CF B	8
Udinese	8
1. FC Magdeburg	8
Konyaspor	8
Salernitana	8
Karagümrük SK	8
Parma	8
Gil Vicente	8
ASSE	8
Aalborg BK	8
Ascoli	8
Brøndby IF	8
Werder Bremen	8
Newell's	8
Henan Jianye	8
Swindon Town	8
Randers FC	8
Mumbai City FC	8
ATK Mohun Bagan FC	8
Tigre	8
Argentinos Jrs.	8
Adelaide United	8
RFC Seraing	8
Pisa	8
FC St. Pauli	8
Deportes Tolima	8
Sheffield Wed	8
Fenerbahçe	8
Deportivo Cali	8
Bolton	8
Málaga CF	8
IFK Göteborg	8
NEC Nijmegen	8
Mansfield Town	7
KV Oostende	7
Wisła Płock	7
Gaziantep	7
SC Austria	7
Adana Demirspor	7
Plymouth Argyle	7
River Plate Montevideo	7
RCD Espanyol	7
Viktoria Köln	7
Junior	7
Lincoln City	7
FC Lorient	7
UD Almería	7
Quevilly Rouen Métropole	7
Uni. Católica	7
Hannover 96	7
Marítimo	7
Charlton Ath	7
Erzgebirge Aue	7
Mjällby AIF	7
Clermont	7
Everton de Viña del Mar	7
Sivasspor	7
Dynamo Kyiv	7
CD Lugo	7

Colchester	7
SV Sandhausen	7
VfB Oldenburg	7
Standard Liège	7
FBC Melgar	7
Jagiellonia	7
Newcastle Jets	7
Shakhtar Donetsk	7
Racing de Santander	7
NY Red Bulls	7
Odense BK	7
Rongcheng FC	7
Club Atlético Platense	7
Dundee United	7
Havre AC	7
Brisbane Roar	7
Pau FC	7
Rayo Vallecano	7
Rio Ave	7
Dinamo Zagreb	7
Drogheda United FC	7
CD Mirandés	7
Derby County	7
FC Vizela	7
FC Volendam	7
Pohang Steelers	7
Tianjin TEDA	7
Beijing Guoan	7
Oriente Petrolero	7
Birmingham City	7
Hyderabad FC	7
Atl. Nacional	7
1. FSV Mainz 05	7
Carlisle United	7
PAOK	7
Chennaiyin FC	7
Well. Phoenix	7
Colo-Colo	7
Morecambe	7
Kerala Blasters FC	7
Degerfors IF	6
KRC Genk	6
HNK Hajduk Split	6
FC Lugano	6
NorthEast United FC	6
Odds BK	6
Bradford City	6
Widzew Łódź	6
MKE Ankaragücü	6
Stade de Reims	6
Finn Harps	6
Trabzonspor	6
FC Ingolstadt	6
Crewe Alexandra	6
SCR Altach	6
Ipswich	6
FK Haugesund	6
SC Verl	6
St. Mirren	6
Motherwell	6

Korona Kielce	6
D. Alavés	6
SpVgg Bayreuth	6
Northampton	6
River Plate	6
Barrow	6
Ross County	6
Portsmouth	6
Odisha FC	6
Benevento	6
Bengaluru FC	6
Jamshedpur FC	6
FC Hermannstadt	6
San Lorenzo	6
Lillestrøm SK	6
FC Annecy	6
Independiente Medellín	6
Patronato	6
Hartlepool United	6
Hatayspor	6
AIK	6
Al Fateh	6
Rosario Central	6
Sparta Rotterdam	6
FC Luzern	6
Hamburger SV	6
Kayserispor	6
Preston	6
Sporting KC	6
SM Caen	6
FC Andorra	6
Holstein Kiel	6
Ferencvárosi TC	6
Olimpia	6
Levante UD	6
Stade Lavallois	6
FC Emmen	6
SD Ponferradina	6
HJK Helsinki	6
SC Bastia	6
GIF Sundsvall	5
UCD	5
Huracán	5
Sporting Cristal	5
Bari	5
Excelsior	5
VAFC	5
Reggina	5
SC East Bengal	5
IFK Värnamo	5
Frosinone	5
Cosenza	5
The Strongest	5
Santa Clara	5
Aldosivi	5
Sint-Truiden	5
Ümraniyespor	5
Central Coast	5
SC Heerenveen	5
Wilstermann	5

```
SD Huesca                          5
Helsingborgs IF                    5
WS Wanderers                       5
Giresunspor                        5
Go Ahead Eagles                    5
Gimnasia                           5
Amiens SC                          5
Defensa                            5
Bielefeld                          5
EA Guingamp                        5
Kaizer Chiefs                      5
Abha Club                          5
Dundalk                            4
Macarthur FC                       4
Fortuna Sittard                    4
Casa Pia AC                        4
Guaireña FC                        4
Unión La Calera                    4
Ayacucho                           4
LDU Quito                          4
Sydney FC                          4
FC U Craiova 1948                  4
IK Sirius                          4
Universidad Católica del Ecuador   3
PGE Stal Mielec                    3
Always Ready                       3
Hallescher FC                      3
Tromsø IL                          3
Portimonense SC                    3
Wolfsberger AC                     3
Metropolitanos                     3
Orlando Pirates                    3
SJ Earthquakes                     3
Barracas Central                   3
Gimcheon Sangmu                    3
Alianza Lima                       3
Lech Poznań                        2
Peñarol                            2
Independiente Petrolero            2
İstanbulspor                       2
Caracas FC                         2
Dep. La Guaira                     2
Barcelona SC                       1
9 de Octubre                       1
Montevideo Wanderers               1
Name: Club, dtype: int64
```

Club

It has many values, I'll create a new column which will have 1 if the club is FUT ICONS or HERO and 0 otherwise.

In [48]:
```
# Defining the map function
def icon_map(x):
    if x == 'FUT ICONS':
        return 1
    elif x == 'HERO':
        return 1
```

```
    else:
        return 0

# Applying the function to the club column
futbin_data['Club_Hero'] = futbin_data['Club'].apply(icon_map)
```

In [49]:
```
futbin_data.Club_Hero.value_counts()
```

Out[49]:
```
0    6231
1     166
Name: Club_Hero, dtype: int64
```

In [50]:
```
futbin_data = futbin_data.drop(['Club'], 1)
```

In [51]:
```
futbin_data.head()
```

Out[51]:

	Name	Nation	League	Rating	Main_Position	Alternate_Positions	Card_Version	Run_Style
0	Pelé	Brazil	Icons	98	CAM	CF,ST	Icon	Explosive
1	Lionel Messi	Argentina	Ligue 1	98	RW	RM	TOTY	Controlled
2	Lionel Messi	Argentina	Ligue 1	98	ST	RM,RW	TOTS	Controlled
3	Karim Benzema	France	LaLiga Santander	97	CF	ST	TOTY	Controlled
4	Kylian Mbappé	France	Ligue 1	97	ST	CF,LW	TOTY	Controlled

League

In [52]:
```
futbin_data.League.value_counts()
```

Out[52]:
```
Premier League                306
LaLiga Santander              295
Major League Soccer           290
EFL Championship (ENG 2)      277
Serie A TIM                   277
Ligue 1                       259
Bundesliga                    254
EFL League One (ENG 3)        234
EFL League Two (ENG 4)        210
CONMEBOL Libertadores         195
LaLiga SmartBank (ESP 2)      190
Chinese FA Super L. (CHN 1)   176
3. Liga (GER 3)               174
CONMEBOL Sudamericana         174
Bundesliga 2 (GER 2)          171
Eredivisie                    169
Italy Serie B (2)             168
MBS Pro League (SAU 1)        164
Liga NOS (POR 1)              162
Polski Ekstraklasa (POL 1)    162
1A Pro League (BEL 1)         161
Liga I (ROM 1)                159
Ligue 2 (FRA 2)               154
Süper Lig (TUR 1)             153
Eliteserien (NOR 1)           134
Allsvenskan (SWE 1)           134
K League 1 (KOR 1)            130
Icons                         128
3F Superliga (DEN 1)          116
Primera División (ARG 1)      110
Ö Bundesliga (AUT 1)          105
Scottish Premiership (SPFL)   102
A-League (AUS 1)               93
Raiffeisen Super L. (SUI 1)    91
SSE Airtricity League (IRL 1)  86
Indian Super League (IND 1)    74
Česká Liga (CZE 1)             29
Hellas Liga (GRE 1)            29
South African FL (RSA 1)       18
Ukrayina Liha (UKR 1)          14
United Emirates L. (UAE 1)     13
Protathlima Cyta (CYP 1)       13
Liga Hrvatska (CRO 1)          13
National League (ENG 5)        12
Liga Dimayor II                 7
Nemzetti Bajnoksag (HUN 1)      6
Finnliiga (FIN 1)               6
Name: League, dtype: int64
```

We have many leagues, I'll collect the 5 mayor european leagues, and some special cases (Worls Cup, Icons), and all of the other values will be other. After this simplification I'll create categorical columns from the values.

In [53]:
```
# Defining the map function
def special_club_map(x):
    if x == 'World Cup':
        return 'World Cup'
    elif x == 'Icons':
        return 'Icons'
```

```
    elif x == 'LaLiga Santander':
        return 'LaLiga Santander'
    elif x == 'Premier League':
        return 'Premier League'
    elif x == 'Major League Soccer':
        return 'Major League Soccer'
    elif x == 'Serie A TIM':
        return 'Serie A TIM'
    elif x == 'Ligue 1':
        return 'Ligue 1'
    elif x == 'Bundesliga':
        return 'Bundesliga'
    else:
        return 'Other'

# Applying the function to the club column
futbin_data['League_Cat'] = futbin_data['League'].apply(special_club_map)
```

In [54]:
```
futbin_data.League_Cat.value_counts()
```

Out[54]:
```
Other                 4588
Premier League         306
LaLiga Santander       295
Major League Soccer    290
Serie A TIM            277
Ligue 1                259
Bundesliga             254
Icons                  128
Name: League_Cat, dtype: int64
```

In [55]:
```
futbin_data = futbin_data.drop(['League'], 1)
```

In [56]:
```
futbin_data.head()
```

Out[56]:

	Name	Nation	Rating	Main_Position	Alternate_Positions	Card_Version	Run_Style	Price
0	Pelé	Brazil	98	CAM	CF,ST	Icon	Explosive	3270000.0
1	Lionel Messi	Argentina	98	RW	RM	TOTY	Controlled	4350000.0
2	Lionel Messi	Argentina	98	ST	RM,RW	TOTS	Controlled	4640000.0
3	Karim Benzema	France	97	CF	ST	TOTY	Controlled	1850000.0
4	Kylian Mbappé	France	97	ST	CF,LW	TOTY	Controlled	9750000.0

Nation

I'll check the nation column, but I have a feeling that I need to remove it, because it has probably a lot of values and it will be hard to extract information from it.

In [57]:
```
futbin_data.Nation.value_counts()
```

Out[57]:

England	645
Germany	424
Spain	395
France	363
Argentina	341
Italy	253
Netherlands	180
Brazil	175
China PR	159
Sweden	148
Portugal	143
Romania	141
Republic of Ireland	140
United States	139
Poland	138
Norway	128
Denmark	126
Korea Republic	123
Belgium	118
Saudi Arabia	114
Austria	99
Australia	95
Scotland	91
Colombia	88
Switzerland	81
Turkey	74
Wales	63
Uruguay	62
Croatia	57
India	57
Paraguay	52
Chile	48
Senegal	46
Nigeria	45
Côte d'Ivoire	39
Ghana	39
Morocco	37
Japan	37
Czech Republic	36
Venezuela	36
Ecuador	32
Mali	32
Canada	31
Greece	30
Finland	30
Serbia	30
Ukraine	29
Cameroon	28
Northern Ireland	27
Bosnia and Herzegovina	26
Peru	24
Mexico	21
Slovakia	21
South Africa	20
Algeria	17
Congo DR	17
Bolivia	17
New Zealand	16
Slovenia	16
Jamaica	15

Guinea	15
Albania	14
Hungary	13
Iceland	13
Kosovo	12
Tunisia	11
Gambia	11
Costa Rica	10
United Arab Emirates	10
Montenegro	9
Iran	9
Guinea-Bissau	9
Bulgaria	9
Congo	9
Burkina Faso	8
Georgia	8
FYR Macedonia	8
Russia	8
Benin	8
Cape Verde Islands	8
Luxembourg	7
Cyprus	7
Gabon	7
Sierra Leone	6
Panama	6
Togo	6
Angola	5
Egypt	5
Comoros	5
Curaçao	4
Zambia	4
Estonia	4
Lithuania	4
Moldova	4
Suriname	4
Haiti	3
Grenada	3
El Salvador	3
Latvia	3
Zimbabwe	3
Honduras	3
Israel	3
Gibraltar	2
Mozambique	2
Madagascar	2
Belarus	2
Iraq	2
Syria	2
Equatorial Guinea	2
Azerbaijan	2
St. Kitts and Nevis	2
Guyana	2
Uzbekistan	2
Liberia	2
Indonesia	1
Hong Kong	1
Papua New Guinea	1
Mauritania	1
Lebanon	1
Uganda	1

```
Philippines                1
Mauritius                  1
Armenia                    1
St. Lucia                  1
Guatemala                  1
Antigua and Barbuda        1
Bermuda                    1
Korea DPR                  1
Libya                      1
Fiji                       1
Namibia                    1
Dominican Republic         1
Name: Nation, dtype: int64
```

We have a few nations more than 1000 players so I decided that I will keep those, and will use other for the others. It is just a subjective decision, so later it can be modified, and maybe keep other nations as well.

One other derived direction could be to merge the countries by continents Europe, America, Asia, and so on...

In [58]:
```
# Defining the map function
def nation_keep_map(x):
    if x == 'England':
        return 'England'
    elif x == 'Germany':
        return 'Germany'
    elif x == 'Spain':
        return 'Spain'
    elif x == 'France':
        return 'France'
    elif x == 'Argentina':
        return 'Argentina'
    else:
        return 'Other'

# Applying the function to the club column
futbin_data['Nation_Cat'] = futbin_data['Nation'].apply(nation_keep_map)
```

In [59]:
```
futbin_data.Nation_Cat.value_counts()
```

Out[59]:
```
Other        4229
England       645
Germany       424
Spain         395
France        363
Argentina     341
Name: Nation_Cat, dtype: int64
```

In [60]:
```
futbin_data = futbin_data.drop(['Nation'], 1)
```

In [61]:
```
futbin_data.head()
```

Out[61]:

	Name	Rating	Main_Position	Alternate_Positions	Card_Version	Run_Style	Price	Skills_Sta
0	Pelé	98	CAM	CF,ST	Icon	Explosive	3270000.0	5
1	Lionel Messi	98	RW	RM	TOTY	Controlled	4350000.0	4
2	Lionel Messi	98	ST	RM,RW	TOTS	Controlled	4640000.0	4
3	Karim Benzema	97	CF	ST	TOTY	Controlled	1850000.0	4
4	Kylian Mbappé	97	ST	CF,LW	TOTY	Controlled	9750000.0	5

Card_Version

In [62]:
```
futbin_data.Card_Version.value_counts()
```

Out[62]:

Normal	4711
non-rare	353
Rare	221
Libertadores	175
IF	165
Sudamericana	139
Icon	92
TOTS	43
SIF	32
CF	26
Hero	22
World Cup ICON	18
CM	16
Fantasy	16
Winter Wildcards	14
LWB	12
Fantasy Hero	12
FUT Centurions	11
Flashback SBC	11
RWB	10
World Cup PTG	9
Out Of Position SBC	8
Trophy Titans - ICON	8
TOTY	8
CDM	8
World Cup Phenoms	8
World Cup Showdown S	7
Winter Wildcards SBC	7
FUT Future Stars	7
UCL Live	7
FUT Birthday	6
OTW	6
MOTM	6
Player Moments SBC	5
Out Of Position	5
FUT Birthday SBC	5
Non-Rare	5
World Cup Star SBC	5
Future Stars Token	5
PL POTM SBC	5
LM,RW	5
RB,RM	4
World Cup TOTT	4
Champions League	4
Dynamic Duos	4
LWB,LM	4
RM	4
Rulebreakers	4
LB	4
UEL Live	4
FUT Centurions SBC	4
ShowDown	4
ShowDown SBC	4
CDM,CAM	3
RM,LW	3
TOTY Honourable Ment	3
TIF	3
Trophy Titans Hero	3
RW	3
Ligue 1 POTM SBC	3

TOTY Icon	3
Bundes POTM SBC	3
FUT BD Token	3
Wildcard Token	3
TOTS Moments	3
TOTS Moments SBC	3
LW	2
CB,CM	2
CM,ST	2
World Cup ICON SBC	2
CAM,RW	2
CAM,CF	2
CAM	2
RB	2
ST,LW	2
ST	2
LM	2
LM,CAM,RW	2
LM,RW,ST	2
CB	2
Europa League	2
CM,CF	2
Fantasy SBC	2
Europa League MOTM	2
World Cup Stories	2
LB,LM	2
FUT Ballers SBC	2
POTM Serie A SBC	2
CAM,ST	1
CAM,LW	1
RWB,LB,RM	1
CB,LB,LM	1
RM,CM	1
UECL Live	1
FUT Future Stars SBC	1
CM,CAM	1
World Cup Hero	1
RB,CB	1
RWB,RB	1
Eredivisie POTM SBC	1
FIF	1
Wildcard Token SBC	1
RB,LB	1
LWB,LM,RW	1
RWB,RM	1
CF,LW	1
RM,CF,LW	1
LM,CF,RW	1
CAM,ST,LW	1
UCL LIVE	1
LB,RM,LM	1
RM,CAM,LW	1
LB,LWB	1
TOTS SBC	1
Conference League	1
Road to World Cup	1
CL MOTM	1
LM,CAM,CF	1
UCL Live SBC	1

```
World Cup Phenoms SB       1
Name: Card_Version, dtype: int64
```

too many options and the values are mixed, not just card versions (toty, fut hero, winter wildcards, etc) but positions (LB, RB, etc and these combinations as well)

I'll just remove this column for now.

In [63]:
```
futbin_data = futbin_data.drop(['Card_Version'], 1)
```

In [64]:
```
futbin_data.head()
```

Out[64]:

	Name	Rating	Main_Position	Alternate_Positions	Run_Style	Price	Skills_Star	Weak_Foot_S
0	Pelé	98	CAM	CF,ST	Explosive	3270000.0	5	
1	Lionel Messi	98	RW	RM	Controlled	4350000.0	4	
2	Lionel Messi	98	ST	RM,RW	Controlled	4640000.0	4	
3	Karim Benzema	97	CF	ST	Controlled	1850000.0	4	
4	Kylian Mbappé	97	ST	CF,LW	Controlled	9750000.0	5	

Positions

In [65]:
```
futbin_data.Main_Position.value_counts()
```

Out[65]:
```
CB     1107
ST      921
CM      740
GK      685
CDM     507
CAM     400
RB      396
LB      392
RM      292
LM      285
RW      192
LW      189
RWB     120
LWB      98
CF       73
Name: Main_Position, dtype: int64
```

would be perfect for categorical columns, will convert it at the last step.

In [66]:
```
futbin_data.Alternate_Positions.value_counts()
```

Out[66]:

0	2028
CF	653
CM	411
CDM	276
LWB	237
RWB	198
RB	152
LB	106
LM,RW	95
RM	93
CAM	91
RW	89
RM,LW	80
LM	77
LWB,LM	73
LW	70
RWB,RM	61
CF,LW	60
LM,CF	53
RWB,LB	50
CDM,CAM	49
CB	42
ST,LW	41
LM,ST	40
RWB,CB	40
ST	40
CF,RW	35
RM,CF	34
RM,LM	33
RB,RM	32
CAM,LW	31
RW,ST	31
CB,CM	31
CB,LWB	29
LM,RW,ST	29
RM,CAM,LW	29
RM,ST	28
CAM,CF	27
LB,LM	26
RM,ST,LW	26
LM,CAM	25
CAM,RW	23
RW,LW	22
CF,RW,LW	21
CDM,CM	21
RM,CM	21
RM,CAM	21
RM,LM,CF	19
CAM,ST	19
RB,LWB	18
CM,RW	18
CM,LM	15
CM,CF	15
CM,LW	14
CB,LB	14
RB,RW	13
RM,RW	13
RB,CM	13
LB,LW	12
LWB,LW	12

```
RM,CM,LW        11
RWB,CM          10
CM,CAM          10
RM,CAM,CF       10
RB,LB           10
LM,CAM,CF       10
RB,CB           10
LB,LWB           9
LM,CAM,RW        9
CF,ST            9
CAM,ST,LW        9
LWB,CDM          9
CM,CAM,RW        8
RWB,CB,CDM       8
LM,CF,RW         8
RB,CDM           8
LM,LW            8
RWB,RW           8
CM,LM,RW         7
RWB,RM,RW        7
CM,ST            7
CDM,LM           7
RB,CB,LWB        7
RWB,LB,RM        7
LWB,RM,LM        6
LB,CM            6
CAM,CF,RW        5
LM,CAM,ST        5
LWB,CM,LM        5
RWB,CDM          5
RW,ST,LW         5
RWB,RB,RW        5
RM,CF,LW         5
RWB,RB           4
CB,CDM           4
CM,CAM,LW        4
RM,CAM,ST        4
RM,CF,RW         4
LWB,RM,LW        4
LWB,CM           4
CDM,RM           4
RB,LM,RW         4
CM,CAM,CF        4
LB,RM,LW         3
LB,RM,LM         3
CM,LM,CAM        3
LWB,LM,LW        3
LB,CDM           3
LM,ST,LW         3
CM,CF,RW         3
CM,CAM,ST        3
RM,RW,LW         3
RWB,CB,LB        3
RM,LM,ST         2
CDM,ST           2
LB,CM,LW         2
RB,LWB,CDM       2
CDM,LW           2
CAM,CF,LW        2
RWB,LM,CAM       2
```

RB,RM,LW	2
RWB,LM,RW	2
RWB,LB,CDM	2
RWB,LWB	2
LWB,RM	2
RB,CM,RW	2
RWB,LM	2
RWB,RM,LM	2
CDM,RW	2
RM,LM,CAM	2
RWB,CB,RM	2
LB,RW	2
LM,CF,LW	2
RB,LW	2
CB,LB,LM	2
RM,CM,CAM	2
RWB,LB,LWB	2
CB,LWB,LM	2
RWB,RM,LW	2
RWB,CM,RW	1
RB,LWB,RM	1
CM,LM,ST	1
RB,CB,CM	1
RM,LM,RW	1
CF,RW,ST	1
LWB,RW	1
CB,LB,CDM	1
LWB,LM,RW	1
CB,LWB,CM	1
RB,RM,RW	1
RB,ST	1
RWB,CAM	1
RWB,RW,ST	1
LB,CAM,LW	1
RWB,RM,CM	1
RWB,CF	1
RWB,ST	1
RWB,CAM,CF	1
RWB,LWB,RW	1
CDM,CF	1
CAM,RW,ST	1
LM,CAM,LW	1
RB,RM,LM	1
RB,RW,ST	1
RM,RW,ST	1
RB,LM,ST	1
LB,CDM,LM	1
CM,LM,LW	1
LWB,CM,LW	1
CB,LWB,CDM	1
RWB,CB,RW	1
RB,CDM,RW	1
RB,CB,RM	1
RM,LM,LW	1
LB,CDM,LW	1
LWB,CDM,LW	1
LB,LWB,LM	1
RM,CF,ST	1
CM,RW,LW	1
RWB,RM,CF	1

```
RB,CF,RW          1
RWB,LB,LM         1
LM,RW,LW          1
CF,ST,LW          1
RB,LWB,LM         1
LWB,CAM           1
Name: Alternate_Positions, dtype: int64
```

In [67]:
```
futbin_data['Alt_Pos_List'] = futbin_data['Alternate_Positions'].str.split(',')
```

In [68]:
```
futbin_data['Alt_Pos_List'].head()
```

Out[68]:
```
0    [CF, ST]
1        [RM]
2    [RM, RW]
3        [ST]
4    [CF, LW]
Name: Alt_Pos_List, dtype: object
```

In [69]:
```
#futbin_data[['Alt_Pos_1', 'Alt_Pos_2', 'Alt_Pos_3']] = futbin_data['Alternate_Positic
```

In [70]:
```
futbin_data = futbin_data.join(futbin_data['Alternate_Positions'].str.split(',', expar
```

In [71]:
```
futbin_data.head()
```

Out[71]:

	Name	Rating	Main_Position	Alternate_Positions	Run_Style	Price	Skills_Star	Weak_Foot_S
0	Pelé	98	CAM	CF,ST	Explosive	3270000.0	5	
1	Lionel Messi	98	RW	RM	Controlled	4350000.0	4	
2	Lionel Messi	98	ST	RM,RW	Controlled	4640000.0	4	
3	Karim Benzema	97	CF	ST	Controlled	1850000.0	4	
4	Kylian Mbappé	97	ST	CF,LW	Controlled	9750000.0	5	

In [72]:
```
futbin_data['Alt_Pos_Count'] = futbin_data['Alt_Pos_List'].str.len()
```

In [73]:
```
futbin_data = futbin_data.drop(['Alternate_Positions'], 1)
```

In [74]:
```
futbin_data = futbin_data.drop(['Alt_Pos_List'], 1)
```

In [75]:
```
futbin_data.head()
```

Out[75]:

	Name	Rating	Main_Position	Run_Style	Price	Skills_Star	Weak_Foot_Star	Attack_Workrat
0	Pelé	98	CAM	Explosive	3270000.0	5	4	H
1	Lionel Messi	98	RW	Controlled	4350000.0	4	4	H
2	Lionel Messi	98	ST	Controlled	4640000.0	4	4	H
3	Karim Benzema	97	CF	Controlled	1850000.0	4	5	H
4	Kylian Mbappé	97	ST	Controlled	9750000.0	5	4	H

In [76]:
```
futbin_data.tail()
```

Out[76]:

	Name	Rating	Main_Position	Run_Style	Price	Skills_Star	Weak_Foot_Star	Attack_Workrate
7154	Hao Ning	48	ST	Controlled	200.0	2	3	M
7155	Jacen Russell-Rowe	48	ST	Controlled	200.0	2	3	M
7156	Andre? Brînzea	48	GK	Controlled	200.0	0	2	M
7158	Junjie Wu	46	LB	Controlled	200.0	2	3	M
7159	Kailin Barlow	46	CDM	Controlled	250.0	0	2	H

In [77]:
```
futbin_data.Alt_Pos_1.replace('0', np.nan, inplace=True)
```

In [78]:
```
futbin_data.Alt_Pos_2.fillna(value=np.nan, inplace=True)
```

In [79]:
```
futbin_data.Alt_Pos_3.fillna(value=np.nan, inplace=True)
```

Need to update Alt_Pos_Count to 0 if the Alt_Pos_1 is NaN, because originally, the field conatins 0 instead of null value.

In [80]:
```
#futbin_data['Alt_Pos_Count'] = futbin_data.apply(lambda x: 0 if x['Alt_Pos_1']=='NaN
futbin_data.Alt_Pos_Count = np.where(futbin_data.Alt_Pos_1.isna(), 0, futbin_data.Alt_
```

In [81]:
```
futbin_data.Alt_Pos_1.info()
```

```
<class 'pandas.core.series.Series'>
Int64Index: 6397 entries, 0 to 7159
Series name: Alt_Pos_1
Non-Null Count  Dtype
--------------  -----
4369 non-null   object
dtypes: object(1)
memory usage: 358.0+ KB
```

In [82]:
```
futbin_data[futbin_data.Alt_Pos_1.isna()].tail()
```

Out[82]:

	Name	Rating	Main_Position	Run_Style	Price	Skills_Star	Weak_Foot_Star	Attack_Workrate
7150	Peinan Li	48	CB	Controlled	200.0	0	3	M
7151	Wenzhe Zhao	48	CB	Controlled	200.0	0	3	M
7152	Xiongtao Deng	48	GK	Controlled	200.0	0	3	M
7156	Andre? Brînzea	48	GK	Controlled	200.0	0	2	M
7159	Kailin Barlow	46	CDM	Controlled	250.0	0	2	H

In [83]:
```
futbin_data.tail()
```

Out[83]:

	Name	Rating	Main_Position	Run_Style	Price	Skills_Star	Weak_Foot_Star	Attack_Workrate
7154	Hao Ning	48	ST	Controlled	200.0	2	3	M
7155	Jacen Russell-Rowe	48	ST	Controlled	200.0	2	3	M
7156	Andre? Brînzea	48	GK	Controlled	200.0	0	2	M
7158	Junjie Wu	46	LB	Controlled	200.0	2	3	M
7159	Kailin Barlow	46	CDM	Controlled	250.0	0	2	H

3. Categorical columns

Main_Position

Run_Style

Attack_Workrate

Defense_Workrate

BodyType_Text

League_Cat

Nation_Cat

==========

Alt_Pos_1

Alt_Pos_2

Alt_Pos_3

First I'll handle the "normal" cat columns, skipping the from alt pos, because those could have emty fields.

In [84]:
```
# Creating a dummy variable for some of the categorical variables and dropping the fir
dummy_fields = pd.get_dummies(futbin_data[['Main_Position', 'Run_Style', 'Attack_Workr
                                           'BodyType_Text', 'League_Cat', 'Nation_Cat'

# Adding the results to the master dataframe
futbin_data = pd.concat([futbin_data, dummy_fields], axis=1)
```

In [85]:
```
futbin_data.head()
```

Out[85]:

	Name	Rating	Main_Position	Run_Style	Price	Skills_Star	Weak_Foot_Star	Attack_Workrat
0	Pelé	98	CAM	Explosive	3270000.0	5	4	H
1	Lionel Messi	98	RW	Controlled	4350000.0	4	4	H
2	Lionel Messi	98	ST	Controlled	4640000.0	4	4	H
3	Karim Benzema	97	CF	Controlled	1850000.0	4	5	H
4	Kylian Mbappé	97	ST	Controlled	9750000.0	5	4	H

In [86]:
```
# dropping the repeated variables
futbin_data = futbin_data.drop(['Main_Position', 'Run_Style', 'Attack_Workrate',
                                'Defense_Workrate', 'BodyType_Text', 'League_Cat', 'Nation_Cat
```

In [87]:
```
futbin_data.head()
```

Out[87]:

	Name	Rating	Price	Skills_Star	Weak_Foot_Star	Pace / Diving	Shooting / Handling	Passing / Kicking	Dribbling / Reflexes	De / S
0	Pelé	98	3270000.0	5	4	95.0	96	93	96	
1	Lionel Messi	98	4350000.0	4	4	93.0	98	97	99	
2	Lionel Messi	98	4640000.0	4	4	94.0	97	96	99	
3	Karim Benzema	97	1850000.0	4	5	92.0	97	90	94	
4	Kylian Mbappé	97	9750000.0	5	4	99.0	96	88	98	

Alt positins filling with missing value, and use it as a category

Alt_Pos_1

In [88]:
```
futbin_data.Alt_Pos_1.isnull().sum()
```

Out[88]:
```
2028
```

In [89]:
```
futbin_data.Alt_Pos_1.fillna('missing', inplace=True)
```

In [90]:
```
futbin_data.Alt_Pos_1.isnull().sum()
```

Out[90]:
```
0
```

In [91]:
```
futbin_data.Alt_Pos_1.value_counts()
```

Out[91]:
```
missing    2028
CF          780
CM          525
RM          444
RWB         435
LM          366
CDM         364
LWB         360
RB          286
CAM         208
LB          176
RW          147
CB          127
ST           81
LW           70
Name: Alt_Pos_1, dtype: int64
```

Alt_Pos_2

In [92]:
```
futbin_data.Alt_Pos_2.isnull().sum()
```

Out[92]:
```
4563
```

In [93]:
```
futbin_data.Alt_Pos_2.fillna('missing', inplace=True)
```

In [94]:
```
futbin_data.Alt_Pos_2.isnull().sum()
```

Out[94]:
```
0
```

In [95]:
```
futbin_data.Alt_Pos_2.value_counts()
```

Out[95]:
```
missing    4563
LW          284
RW          270
LM          206
CAM         196
ST          182
CF          162
RM          132
CM          130
LB           89
CB           73
LWB          68
CDM          33
RB            9
Name: Alt_Pos_2, dtype: int64
```

Alt_Pos_3

In [96]:
```
futbin_data.Alt_Pos_3.isnull().sum()
```

Out[96]:
```
6009
```

In [97]:
```
futbin_data.Alt_Pos_3.fillna('missing', inplace=True)
```

In [98]:
```
futbin_data.Alt_Pos_3.isnull().sum()
```

Out[98]:
```
0
```

In [99]:
```
futbin_data.Alt_Pos_3.value_counts()
```

Out[99]:
```
missing    6009
LW          146
RW           72
ST           51
CF           45
LM           25
CDM          14
RM           11
CAM           9
LWB           9
LB            3
CM            3
Name: Alt_Pos_3, dtype: int64
```

In [100…
```
futbin_data.head()
```

Out[100]:

	Name	Rating	Price	Skills_Star	Weak_Foot_Star	Pace / Diving	Shooting / Handling	Passing / Kicking	Dribbling / Reflexes	De / S
0	Pelé	98	3270000.0	5	4	95.0	96	93	96	
1	Lionel Messi	98	4350000.0	4	4	93.0	98	97	99	
2	Lionel Messi	98	4640000.0	4	4	94.0	97	96	99	
3	Karim Benzema	97	1850000.0	4	5	92.0	97	90	94	
4	Kylian Mbappé	97	9750000.0	5	4	99.0	96	88	98	

Create categorical columns for the alt pos columns

In [101…
```
# Creating a dummy variable for the alt_pos categorical variables and dropping the fir
dummy_altpos = pd.get_dummies(futbin_data[['Alt_Pos_1', 'Alt_Pos_2', 'Alt_Pos_3']], dr

# Adding the results to the master dataframe
futbin_data = pd.concat([futbin_data, dummy_altpos], axis=1)
```

In [102…
```
futbin_data.head()
```

Out[102]:

	Name	Rating	Price	Skills_Star	Weak_Foot_Star	Pace / Diving	Shooting / Handling	Passing / Kicking	Dribbling / Reflexes	De / S
0	Pelé	98	3270000.0	5	4	95.0	96	93	96	
1	Lionel Messi	98	4350000.0	4	4	93.0	98	97	99	
2	Lionel Messi	98	4640000.0	4	4	94.0	97	96	99	
3	Karim Benzema	97	1850000.0	4	5	92.0	97	90	94	
4	Kylian Mbappé	97	9750000.0	5	4	99.0	96	88	98	

In [103…
```
# dropping the repeated variables
futbin_data = futbin_data.drop(['Alt_Pos_1', 'Alt_Pos_2', 'Alt_Pos_3'], 1)
```

In [104…
```
futbin_data.head()
```

Out[104]:

	Name	Rating	Price	Skills_Star	Weak_Foot_Star	Pace / Diving	Shooting / Handling	Passing / Kicking	Dribbling / Reflexes	De / S
0	Pelé	98	3270000.0	5	4	95.0	96	93	96	
1	Lionel Messi	98	4350000.0	4	4	93.0	98	97	99	
2	Lionel Messi	98	4640000.0	4	4	94.0	97	96	99	
3	Karim Benzema	97	1850000.0	4	5	92.0	97	90	94	
4	Kylian Mbappé	97	9750000.0	5	4	99.0	96	88	98	

4. Export final dataset

In [105…

```
csv_futbin = futbin_data.to_csv('futbin.csv', index = False)
print('\nCSV String:\n', csv_futbin)
```

```
CSV String:
 None
```

In [ ]:

In [ ]: