

# Part 2 : Football Transfer Market Data Preprocessing

This notebook covers data cleaning, feature engineering, label encoding, scaling and general preprocessing for modelling.

## Importing Dependencies

In [219...]

```
# Visualization Libraries
import matplotlib
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import mglearn

#Preprocessing Libraries
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.metrics import precision_score, recall_score, confusion_matrix, classification_report
from sklearn import preprocessing
from sklearn.preprocessing import MinMaxScaler, RobustScaler

# ML Libraries
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.decomposition import PCA

# Evaluation Metrics
from yellowbrick.classifier import ClassificationReport
from sklearn import metrics

import warnings# Import this library to suppress the warnings
warnings.filterwarnings('ignore') # The object 'warnings' is used to call the method
```

## Loading data

In [4]:

```
# Reading in the data and assign it to 'df' variable/object
df = pd.read_csv("Merged_Data_4_Pre_Processing.csv")
```

## Confirming load and viewing general information on dataset

In [5]:

```
df.shape
```

Out[5]:

```
(30130, 79)
```

In [6]:

```
df.info()
```

```

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 30130 entries, 0 to 30129
Data columns (total 79 columns):
 #   Column           Non-Null Count Dtype  
 --- 
 0   player_id        30130 non-null  int64   
 1   first_name       28173 non-null  object  
 2   last_name        30130 non-null  object  
 3   name              30130 non-null  object  
 4   last_season      30130 non-null  int64   
 5   current_club_id  30130 non-null  int64   
 6   player_code       30130 non-null  object  
 7   country_of_birth 27528 non-null  object  
 8   city_of_birth     28013 non-null  object  
 9   country_of_citizenship 29587 non-null  object  
 10  date_of_birth    30130 non-null  object  
 11  sub_position     30130 non-null  object  
 12  position          30130 non-null  object  
 13  foot              27844 non-null  object  
 14  height_in_cm     28099 non-null  float64 
 15  market_value_in_eur 30130 non-null  float64 
 16  highest_market_value_in_eur 30130 non-null  float64 
 17  contract_expiration_date 18799 non-null  object  
 18  image_url         30130 non-null  object  
 19  url               30130 non-null  object  
 20  current_club_domestic_competition_id 30130 non-null  object  
 21  current_club_name 30130 non-null  object  
 22  age                30130 non-null  int64   
 23  remaining_contract_days 18799 non-null  float64 
 24  age_group         30130 non-null  object  
 25  games_total       30130 non-null  int64   
 26  goals_total       30130 non-null  int64   
 27  assists_total     30130 non-null  int64   
 28  minutes_played_total 30130 non-null  int64   
 29  goals_for_total   30130 non-null  int64   
 30  goals_against_total 30130 non-null  int64   
 31  clean_sheet_total 30130 non-null  int64   
 32  yellow_cards_total 30130 non-null  int64   
 33  red_cards_total   30130 non-null  int64   
 34  games_2019         30130 non-null  float64 
 35  goals_2019         30130 non-null  float64 
 36  assists_2019       30130 non-null  float64 
 37  minutes_played_2019 30130 non-null  float64 
 38  goals_for_2019     30130 non-null  float64 
 39  goals_against_2019 30130 non-null  float64 
 40  clean_sheet_2019   30130 non-null  float64 
 41  yellow_cards_2019  30130 non-null  float64 
 42  red_cards_2019     30130 non-null  float64 
 43  games_2020         30130 non-null  float64 
 44  goals_2020         30130 non-null  float64 
 45  assists_2020        30130 non-null  float64 
 46  minutes_played_2020 30130 non-null  float64 
 47  goals_for_2020     30130 non-null  float64 
 48  goals_against_2020 30130 non-null  float64 
 49  clean_sheet_2020   30130 non-null  float64 
 50  yellow_cards_2020  30130 non-null  float64 
 51  red_cards_2020     30130 non-null  float64 
 52  games_2021         30130 non-null  float64 
 53  goals_2021         30130 non-null  float64 
 54  assists_2021        30130 non-null  float64 
 55  minutes_played_2021 30130 non-null  float64 
 56  goals_for_2021     30130 non-null  float64 

```

```
59 yellow_cards_2021           30130 non-null float64
60 red_cards_2021              30130 non-null float64
61 games_2022                  30130 non-null float64
62 goals_2022                  30130 non-null int64
63 assists_2022                30130 non-null int64
64 minutes_played_2022         30130 non-null float64
65 goals_for_2022              30130 non-null float64
66 goals_against_2022          30130 non-null float64
67 clean_sheet_2022             30130 non-null int64
68 yellow_cards_2022            30130 non-null int64
69 red_cards_2022               30130 non-null int64
70 games_2023                  30130 non-null float64
71 goals_2023                  30130 non-null float64
72 assists_2023                30130 non-null float64
73 minutes_played_2023          30130 non-null float64
74 goals_for_2023              30130 non-null float64
75 goals_against_2023           30130 non-null float64
76 clean_sheet_2023             30130 non-null float64
77 yellow_cards_2023            30130 non-null float64
78 red_cards_2023               30130 non-null float64
dtypes: float64(44), int64(18), object(17)
memory usage: 18.2+ MB
```

## Generating a profile report for df

```
In [7]: # from ydata_profiling import ProfileReport #importing profile report attribute
# ProfileReport(df)# generating a profile report of the dataset
Summarize dataset:  0%|          | 0/5 [00:00<?, ?it/s]
Generate report structure:  0%|          | 0/1 [00:00<?, ?it/s]
Render HTML:  0%|          | 0/1 [00:00<?, ?it/s]
```

# Overview

## Dataset statistics

|                                      |          |
|--------------------------------------|----------|
| <b>Number of variables</b>           | 79       |
| <b>Number of observations</b>        | 30130    |
| <b>Missing cells</b>                 | 34198    |
| <b>Missing cells (%)</b>             | 1.4%     |
| <b>Duplicate rows</b>                | 0        |
| <b>Duplicate rows (%)</b>            | 0.0%     |
| <b>Total size in memory</b>          | 18.2 MiB |
| <b>Average record size in memory</b> | 632.0 B  |

## Variable types

|                    |    |
|--------------------|----|
| <b>Numeric</b>     | 57 |
| <b>Text</b>        | 10 |
| <b>DateTime</b>    | 2  |
| <b>Categorical</b> | 10 |

## Alerts

player\_id is highly overall correlated with  
last\_season and 2 other fields (last\_season, age)

High correlation

Out[7]:

## Feature Engineering

### Removing Unrequired Features

We discovered through EDA that the dataset contains players that have since retired. There entries are innaccurate and may introduce noise and bias. We will thus remove any player entries for players who have retired since 2022.

```
In [8]: df2 = df[df['last_season'] >= 2022]
```

```
In [9]: df2.shape
```

```
Out[9]: (10169, 79)
```

```
In [10]: pd.set_option('display.max_row', None)
pd.set_option('display.max_columns', None)
```

```
In [11]: df2.nunique()
```

```

Out[11]: player_id           10169
          first_name        3421
          last_name         8445
          name              10072
          last_season        2
          current_club_id    274
          player_code        10067
          country_of_birth   152
          city_of_birth       4000
          country_of_citizenship 144
          date_of_birth      6562
          sub_position       13
          position            4
          foot                3
          height_in_cm        46
          market_value_in_eur 134
          highest_market_value_in_eur 159
          contract_expiration_date 48
          image_url           9066
          url                10169
          current_club_domestic_competition_id 14
          current_club_name    274
          age                 28
          remaining_contract_days 48
          age_group           5
          games_total          205
          goals_total          91
          assists_total         63
          minutes_played_total 5227
          goals_for_total      410
          goals_against_total  256
          clean_sheet_total    92
          yellow_cards_total   55
          red_cards_total       6
          games_2019             57
          goals_2019             35
          assists_2019            24
          minutes_played_2019   2203
          goals_for_2019          128
          goals_against_2019     83
          clean_sheet_2019        27
          yellow_cards_2019       21
          red_cards_2019            3
          games_2020             56
          goals_2020             38
          assists_2020            22
          minutes_played_2020   2372
          goals_for_2020          125
          goals_against_2020     86
          clean_sheet_2020        31
          yellow_cards_2020       18
          red_cards_2020            3
          games_2021             55
          goals_2021             35
          assists_2021            23
          minutes_played_2021   2610
          goals_for_2021          130
          goals_against_2021     84
          clean_sheet_2021        27
          yellow_cards_2021       19
          red_cards_2021            3
          games_2022             58

```

```

minutes_played_2022           2839
goals_for_2022                126
goals_against_2022              82
clean_sheet_2022                  29
yellow_cards_2022                  20
red_cards_2022                      4
games_2023                         21
goals_2023                           15
assists_2023                          10
minutes_played_2023            1118
goals_for_2023                     51
goals_against_2023                   36
clean_sheet_2023                     11
yellow_cards_2023                      8
red_cards_2023                        3
dtype: int64

```

In [12]: `df2.columns`

```

Out[12]: Index(['player_id', 'first_name', 'last_name', 'name', 'last_season',
       'current_club_id', 'player_code', 'country_of_birth', 'city_of_birth',
       'country_of_citizenship', 'date_of_birth', 'sub_position', 'position',
       'foot', 'height_in_cm', 'market_value_in_eur',
       'highest_market_value_in_eur', 'contract_expiration_date', 'image_url',
       'url', 'current_club_domestic_competition_id', 'current_club_name',
       'age', 'remaining_contract_days', 'age_group', 'games_total',
       'goals_total', 'assists_total', 'minutes_played_total',
       'goals_for_total', 'goals_against_total', 'clean_sheet_total',
       'yellow_cards_total', 'red_cards_total', 'games_2019', 'goals_2019',
       'assists_2019', 'minutes_played_2019', 'goals_for_2019',
       'goals_against_2019', 'clean_sheet_2019', 'yellow_cards_2019',
       'red_cards_2019', 'games_2020', 'goals_2020', 'assists_2020',
       'minutes_played_2020', 'goals_for_2020', 'goals_against_2020',
       'clean_sheet_2020', 'yellow_cards_2020', 'red_cards_2020', 'games_2021',
       'goals_2021', 'assists_2021', 'minutes_played_2021', 'goals_for_2021',
       'goals_against_2021', 'clean_sheet_2021', 'yellow_cards_2021',
       'red_cards_2021', 'games_2022', 'goals_2022', 'assists_2022',
       'minutes_played_2022', 'goals_for_2022', 'goals_against_2022',
       'clean_sheet_2022', 'yellow_cards_2022', 'red_cards_2022', 'games_2023',
       'goals_2023', 'assists_2023', 'minutes_played_2023', 'goals_for_2023',
       'goals_against_2023', 'clean_sheet_2023', 'yellow_cards_2023',
       'red_cards_2023'],
      dtype='object')

```

Removing features that will not aid in evaluating player market values such as unique identifiers, names and images, aswell as demographical features that may introduce bias. 'highest\_market\_value\_in\_eur' will also be removed as it derived from 'market\_value\_in\_eur', our target label.

In [135...]: `df3 = df2.drop(['player_id', 'first_name', 'last_name', 'name', 'player_code', 'co`

In [136...]: `df3.shape`

Out[136]: (10169, 66)

In [137...]: `df3.info()`

```

<class 'pandas.core.frame.DataFrame'>
Index: 10169 entries, 72 to 30129
Data columns (total 66 columns):
 #   Column           Non-Null Count Dtype  
 --- 
 0   last_season      10169 non-null  int64   
 1   current_club_id  10169 non-null  int64   
 2   sub_position     10169 non-null  object  
 3   position         10169 non-null  object  
 4   foot              9574 non-null  object  
 5   height_in_cm    9518 non-null  float64 
 6   market_value_in_eur 10169 non-null  float64 
 7   current_club_domestic_competition_id 10169 non-null  object  
 8   current_club_name 10169 non-null  object  
 9   age               10169 non-null  int64   
10  remaining_contract_days 8945 non-null  float64 
11  age_group        10169 non-null  object  
12  games_total      10169 non-null  int64   
13  goals_total      10169 non-null  int64   
14  assists_total    10169 non-null  int64   
15  minutes_played_total 10169 non-null  int64   
16  goals_for_total  10169 non-null  int64   
17  goals_against_total 10169 non-null  int64   
18  clean_sheet_total 10169 non-null  int64   
19  yellow_cards_total 10169 non-null  int64   
20  red_cards_total  10169 non-null  int64   
21  games_2019       10169 non-null  float64 
22  goals_2019       10169 non-null  float64 
23  assists_2019     10169 non-null  float64 
24  minutes_played_2019 10169 non-null  float64 
25  goals_for_2019   10169 non-null  float64 
26  goals_against_2019 10169 non-null  float64 
27  clean_sheet_2019 10169 non-null  float64 
28  yellow_cards_2019 10169 non-null  float64 
29  red_cards_2019   10169 non-null  float64 
30  games_2020       10169 non-null  float64 
31  goals_2020       10169 non-null  float64 
32  assists_2020     10169 non-null  float64 
33  minutes_played_2020 10169 non-null  float64 
34  goals_for_2020   10169 non-null  float64 
35  goals_against_2020 10169 non-null  float64 
36  clean_sheet_2020 10169 non-null  float64 
37  yellow_cards_2020 10169 non-null  float64 
38  red_cards_2020   10169 non-null  float64 
39  games_2021       10169 non-null  float64 
40  goals_2021       10169 non-null  float64 
41  assists_2021     10169 non-null  float64 
42  minutes_played_2021 10169 non-null  float64 
43  goals_for_2021   10169 non-null  float64 
44  goals_against_2021 10169 non-null  float64 
45  clean_sheet_2021 10169 non-null  float64 
46  yellow_cards_2021 10169 non-null  float64 
47  red_cards_2021   10169 non-null  float64 
48  games_2022       10169 non-null  float64 
49  goals_2022       10169 non-null  int64   
50  assists_2022     10169 non-null  int64   
51  minutes_played_2022 10169 non-null  float64 
52  goals_for_2022   10169 non-null  float64 
53  goals_against_2022 10169 non-null  float64 
54  clean_sheet_2022 10169 non-null  int64   
55  yellow_cards_2022 10169 non-null  int64   
56  red_cards_2022   10169 non-null  int64   

```

```
59 assists_2023                      10169 non-null float64
60 minutes_played_2023                 10169 non-null float64
61 goals_for_2023                     10169 non-null float64
62 goals_against_2023                  10169 non-null float64
63 clean_sheet_2023                    10169 non-null float64
64 yellow_cards_2023                   10169 non-null float64
65 red_cards_2023                      10169 non-null float64
dtypes: float64(43), int64(17), object(6)
memory usage: 5.2+ MB
```

In [138... df3.head(50)

|     | last_season | current_club_id | sub_position       | position   | foot  | height_in_cm | market_value_in_eu |
|-----|-------------|-----------------|--------------------|------------|-------|--------------|--------------------|
| 72  | 2022        | 418             | Centre-Forward     | Attack     | right | 185.0        | 2500000.           |
| 88  | 2023        | 678             | Goalkeeper         | Goalkeeper | right | 188.0        | 200000.            |
| 117 | 2023        | 367             | Attacking Midfield | Midfield   | right | 180.0        | 2500000.           |
| 132 | 2022        | 1095            | Goalkeeper         | Goalkeeper | right | 184.0        | 200000.            |
| 135 | 2022        | 1519            | Centre-Forward     | Attack     | left  | 185.0        | 200000.            |
| 152 | 2023        | 2425            | Defensive Midfield | Midfield   | right | 182.0        | 100000.            |
| 161 | 2023        | 371             | Goalkeeper         | Goalkeeper | right | 196.0        | 1500000.           |
| 163 | 2022        | 940             | Goalkeeper         | Goalkeeper | left  | 187.0        | 300000.            |
| 175 | 2023        | 58              | Centre-Back        | Defender   | left  | 189.0        | 1500000.           |
| 178 | 2023        | 511             | Goalkeeper         | Goalkeeper | Nan   | 190.0        | 150000.            |
| 181 | 2023        | 336             | Goalkeeper         | Goalkeeper | left  | 190.0        | 1000000.           |
| 186 | 2023        | 380             | Centre-Back        | Defender   | right | 195.0        | 400000.            |
| 187 | 2023        | 294             | Right Winger       | Attack     | left  | 180.0        | 8000000.           |
| 190 | 2022        | 2239            | Centre-Forward     | Attack     | right | 187.0        | 350000.            |
| 191 | 2022        | 703             | Left Winger        | Attack     | left  | 175.0        | 1800000.           |
| 200 | 2023        | 903             | Centre-Forward     | Attack     | right | 180.0        | 400000.            |
| 217 | 2022        | 931             | Centre-Forward     | Attack     | right | 189.0        | 28000000.          |
| 228 | 2023        | 105             | Left-Back          | Defender   | left  | 172.0        | 600000.            |
| 232 | 2023        | 2919            | Centre-Back        | Defender   | right | 180.0        | 1500000.           |
| 238 | 2023        | 347             | Goalkeeper         | Goalkeeper | right | 184.0        | 500000.            |
| 240 | 2022        | 533             | Defensive Midfield | Midfield   | right | 180.0        | 1000000.           |
| 246 | 2023        | 48332           | Central Midfield   | Midfield   | right | 182.0        | 150000.            |
|     |             |                 | Back               | Defender   | left  | 180.0        | 40000000.          |

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

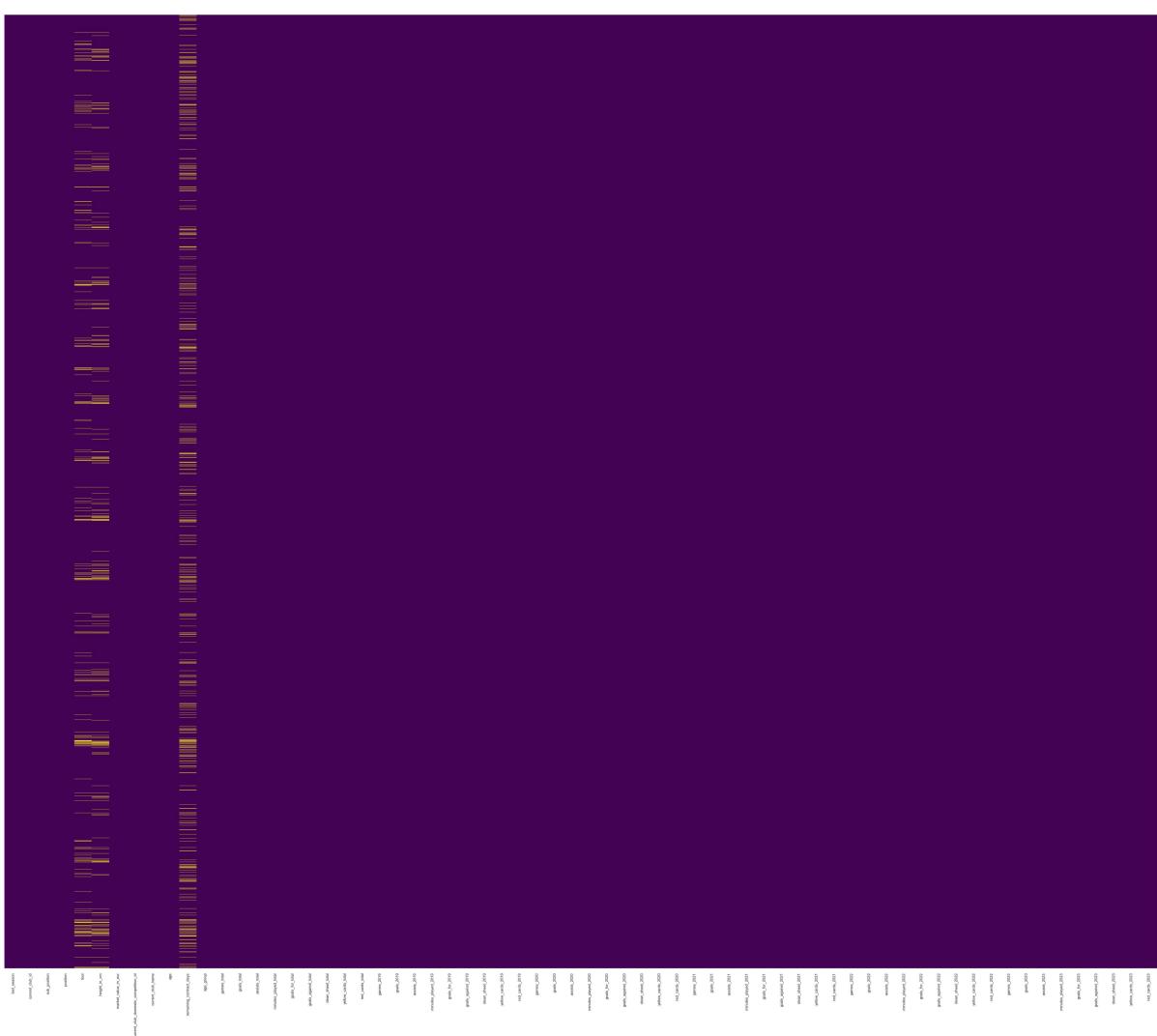
| last_season      | current_club_id | sub_position | position           | foot       | height_in_cm | market_value_in_eu |           |
|------------------|-----------------|--------------|--------------------|------------|--------------|--------------------|-----------|
| 255              | 2023            | 234          | Centre-Back        | Defender   | right        | 183.0              | 5000000.  |
| 258              | 2023            | 232          | Right Midfield     | Midfield   | right        | 177.0              | 2000000.  |
| 261              | 2023            | 3385         | Centre-Back        | Defender   | right        | 190.0              | 100000.   |
| 262              | 2022            | 667          | Centre-Back        | Defender   | left         | 185.0              | 522411.   |
| 277              | 2022            | 1025         | Attacking Midfield | Midfield   | right        | 182.0              | 2500000.  |
| 281              | 2023            | 12321        | Centre-Back        | Defender   | right        | 192.0              | 1500000.  |
| 283              | 2023            | 398          | Right Winger       | Attack     | both         | 167.0              | 1500000.  |
| 287              | 2022            | 2687         | Centre-Back        | Defender   | right        | 186.0              | 600000.   |
| 294              | 2023            | 1894         | Attacking Midfield | Midfield   | right        | 172.0              | 150000.   |
| 295              | 2023            | 8024         | Central Midfield   | Midfield   | left         | 183.0              | 250000.   |
| 302              | 2022            | 678          | Left-Back          | Defender   | both         | 188.0              | 522411.   |
| 313              | 2023            | 273          | Defensive Midfield | Midfield   | left         | 194.0              | 4000000.  |
| 319              | 2022            | 667          | Goalkeeper         | Goalkeeper | right        | 185.0              | 150000.   |
| 323              | 2023            | 3948         | Goalkeeper         | Goalkeeper | left         | 186.0              | 1500000.  |
| 326              | 2023            | 2578         | Central Midfield   | Midfield   | right        | 179.0              | 350000.   |
| 329              | 2023            | 1124         | Right Winger       | Attack     | right        | 175.0              | 350000.   |
| 332              | 2022            | 33           | Centre-Back        | Defender   | right        | 189.0              | 900000.   |
| 336              | 2023            | 873          | Right-Back         | Defender   | right        | 175.0              | 2000000.  |
| 339              | 2023            | 1083         | Attacking Midfield | Midfield   | right        | 177.0              | 800000.   |
| 340              | 2023            | 418          | Centre-Back        | Defender   | right        | 190.0              | 32000000. |
| 344              | 2022            | 683          | Attacking Midfield | Midfield   | left         | 181.0              | 6000000.  |
| 349              | 2023            | 1245         | Centre-Forward     | Attack     | right        | 184.0              | 350000.   |
| Central Midfield |                 |              |                    | Midfield   | left         | 175.0              | 400000.   |

|     | last_season | current_club_id | sub_position       | position | foot  | height_in_cm | market_value_in_eu |
|-----|-------------|-----------------|--------------------|----------|-------|--------------|--------------------|
| 359 | 2023        | 58              | Defensive Midfield | Midfield | right | 178.0        | 5000000.           |
| 370 | 2023        | 60551           | Right Midfield     | Midfield | both  | 166.0        | 600000.            |
| 372 | 2022        | 1096            | Left Winger        | Attack   | right | 181.0        | 2519384.           |
| 383 | 2022        | 3205            | Central            | Midfield | left  | 179.0        | 1300000.           |

## Imputing Missing Values

In [139]: `sns.heatmap(df3.isnull(),yticklabels=False,cbar=False,cmap='viridis')`

Out[139]: <Axes: >



In [140]: `df3.isnull().sum()`

```
Out[140]: last_season          0
           current_club_id      0
           sub_position         0
           position             0
           foot                 595
           height_in_cm         651
           market_value_in_eur   0
           current_club Domestic_competition_id 0
           current_club_name     0
           age                  0
           remaining_contract_days 1224
           age_group            0
           games_total          0
           goals_total          0
           assists_total         0
           minutes_played_total 0
           goals_for_total       0
           goals_against_total   0
           clean_sheet_total     0
           yellow_cards_total    0
           red_cards_total        0
           games_2019             0
           goals_2019              0
           assists_2019            0
           minutes_played_2019    0
           goals_for_2019          0
           goals_against_2019      0
           clean_sheet_2019         0
           yellow_cards_2019        0
           red_cards_2019           0
           games_2020               0
           goals_2020                0
           assists_2020              0
           minutes_played_2020      0
           goals_for_2020            0
           goals_against_2020        0
           clean_sheet_2020           0
           yellow_cards_2020          0
           red_cards_2020             0
           games_2021               0
           goals_2021                0
           assists_2021              0
           minutes_played_2021        0
           goals_for_2021            0
           goals_against_2021          0
           clean_sheet_2021           0
           yellow_cards_2021          0
           red_cards_2021             0
           games_2022               0
           goals_2022                0
           assists_2022              0
           minutes_played_2022        0
           goals_for_2022            0
           goals_against_2022          0
           clean_sheet_2022           0
           yellow_cards_2022          0
           red_cards_2022             0
           games_2023               0
           goals_2023                0
           assists_2023              0
           minutes_played_2023        0
           goals_for_2023            0
                                         0
                                         0
```

```
yellow_cards_2023          0
red_cards_2023             0
dtype: int64
```

```
In [141]: df3['height_in_cm'].describe()
```

```
Out[141]: count    9518.000000
mean     182.666527
std      6.928958
min     160.000000
25%    178.000000
50%    183.000000
75%    188.000000
max     206.000000
Name: height_in_cm, dtype: float64
```

```
In [142... # Calculating the mean market values based on position and age category
mean_height_values = df3.groupby(['position', 'sub_position'])['height_in_cm'].mean()

# Displaying the mean values
print(mean_height_values)
```

|    | position   | sub_position       | height_in_cm |
|----|------------|--------------------|--------------|
| 0  | Attack     | Centre-Forward     | 184.554187   |
| 1  | Attack     | Left Winger        | 176.814056   |
| 2  | Attack     | Right Winger       | 177.224924   |
| 3  | Attack     | Second Striker     | 178.676471   |
| 4  | Defender   | Centre-Back        | 187.966463   |
| 5  | Defender   | Left-Back          | 179.100575   |
| 6  | Defender   | Right-Back         | 179.516129   |
| 7  | Goalkeeper | Goalkeeper         | 190.461938   |
| 8  | Midfield   | Attacking Midfield | 177.996979   |
| 9  | Midfield   | Central Midfield   | 179.848057   |
| 10 | Midfield   | Defensive Midfield | 182.069987   |
| 11 | Midfield   | Left Midfield      | 178.350649   |
| 12 | Midfield   | Right Midfield     | 177.253333   |

```
In [143... # Writing a function of conditional statements for imputing the mean height values
def impute_height(cols):
    height_in_cm = cols[0]
    position = cols[1]
    sub_position = cols[2]

    if pd.isnull(height_in_cm):

        if position == 'Attack':
            if sub_position == 'Centre-Forward':
                return 184.554187
            elif sub_position == 'Left Winger':
                return 176.814056
            elif sub_position == 'Right Winger':
                return 177.224924
            elif sub_position == 'Second Striker':
                return 178.676471
            else:
                return 182.666527

        elif position == 'Defender':
            if sub_position == 'Centre-Back':
                return 187.966463
            elif sub_position == 'Left-Back':
                return 179.100575
            elif sub_position == 'Right-Back':
                return 179.516129
            else:
                return 182.666527

        elif position == 'Goalkeeper':
            return 190.461938

        elif position == 'Midfield':
            if sub_position == 'Attacking Midfield':
                return 177.996979
            elif sub_position == 'Central Midfield':
                return 179.848057
            elif sub_position == 'Defensive Midfield':
                return 182.069987
            elif sub_position == 'Left Midfield':
                return 178.350649
            elif sub_position == 'Right Midfield':
                return 177.253333
            else:
                return 182.666527

    else:
        return height_in_cm
```

```

        return 182.666527

    elif position == 'Goalkeeper':
        if sub_position == 'Goalkeeper':
            return 190.461938
        else:
            return 182.666527

    elif position == 'Midfield':
        if sub_position == 'Attacking Midfield':
            return 177.996979
        elif sub_position == 'Central Midfield':
            return 179.848057
        elif sub_position == 'Defensive Midfield':
            return 182.069987
        elif sub_position == 'Left Midfield':
            return 178.350649
        elif sub_position == 'Right Midfield':
            return 177.253333
        else:
            return 182.666527

    else:
        return 182.666527

else:
    return height_in_cm

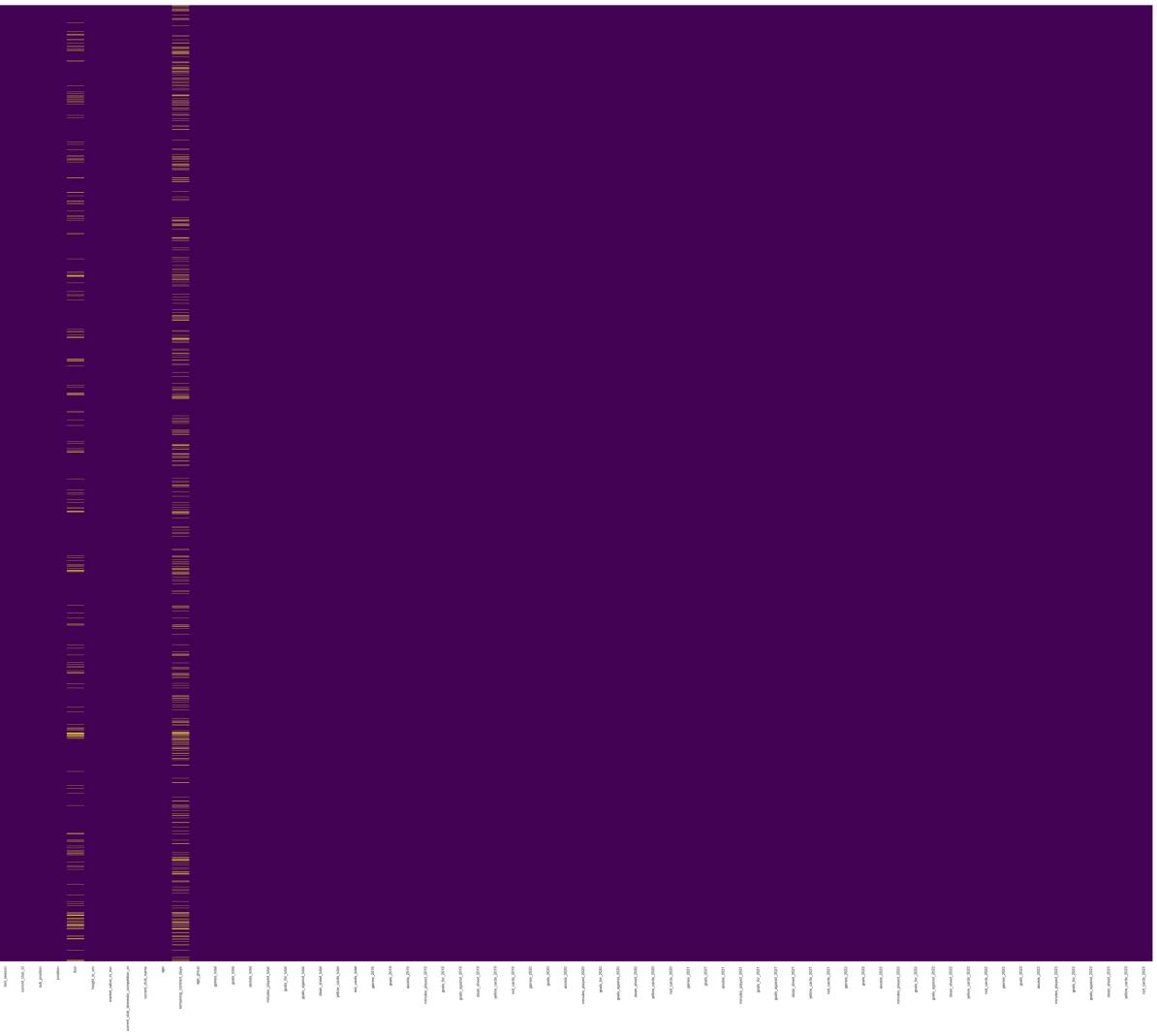
```

In [144]: df4 = df3.copy()

In [145]: # Applying the impute\_height function to the dataframe  
df4['height\_in\_cm'] = df4[['height\_in\_cm', 'position', 'sub\_position']].apply(imput

In [146]: sns.heatmap(df4.isnull(), yticklabels=False, cbar=False, cmap='viridis')

Out[146]: <Axes: >



```
In [147]: df3['foot'].describe()
```

```
Out[147]: count    9574
unique      3
top     right
freq    6737
Name: foot, dtype: object
```

```
In [148]: df4['foot'].unique()
```

```
Out[148]: array(['right', 'left', nan, 'both'], dtype=object)
```

```
In [149... # Calculating the mode for foot values based on position and sub_position
mode_foot_values = df4.groupby(['position', 'sub_position'])['foot'].agg(lambda x:
# Displaying the mode values
print(mode_foot_values)
```

|    | position   | sub_position       | foot  |
|----|------------|--------------------|-------|
| 0  | Attack     | Centre-Forward     | right |
| 1  | Attack     | Left Winger        | right |
| 2  | Attack     | Right Winger       | right |
| 3  | Attack     | Second Striker     | right |
| 4  | Defender   | Centre-Back        | right |
| 5  | Defender   | Left-Back          | left  |
| 6  | Defender   | Right-Back         | right |
| 7  | Goalkeeper | Goalkeeper         | right |
| 8  | Midfield   | Attacking Midfield | right |
| 9  | Midfield   | Central Midfield   | right |
| 10 | Midfield   | Defensive Midfield | right |
| 11 | Midfield   | Left Midfield      | left  |
| 12 | Midfield   | Right Midfield     | right |

In [150]: # Writing a function of conditional statements for imputing the mode for foot based

```
def impute_foot(cols):
    foot = cols[0]
    position = cols[1]
    sub_position = cols[2]

    if pd.isnull(foot):
        if position == 'Attack':
            return 'right'

        elif position == 'Defender':
            if sub_position == 'Left-Back':
                return 'left'
            else:
                return 'right'

        elif position == 'Goalkeeper':
            return 'right'

        elif position == 'Midfield':
            if sub_position == 'Left Midfield':
                return 'left'
            else:
                return 'right'

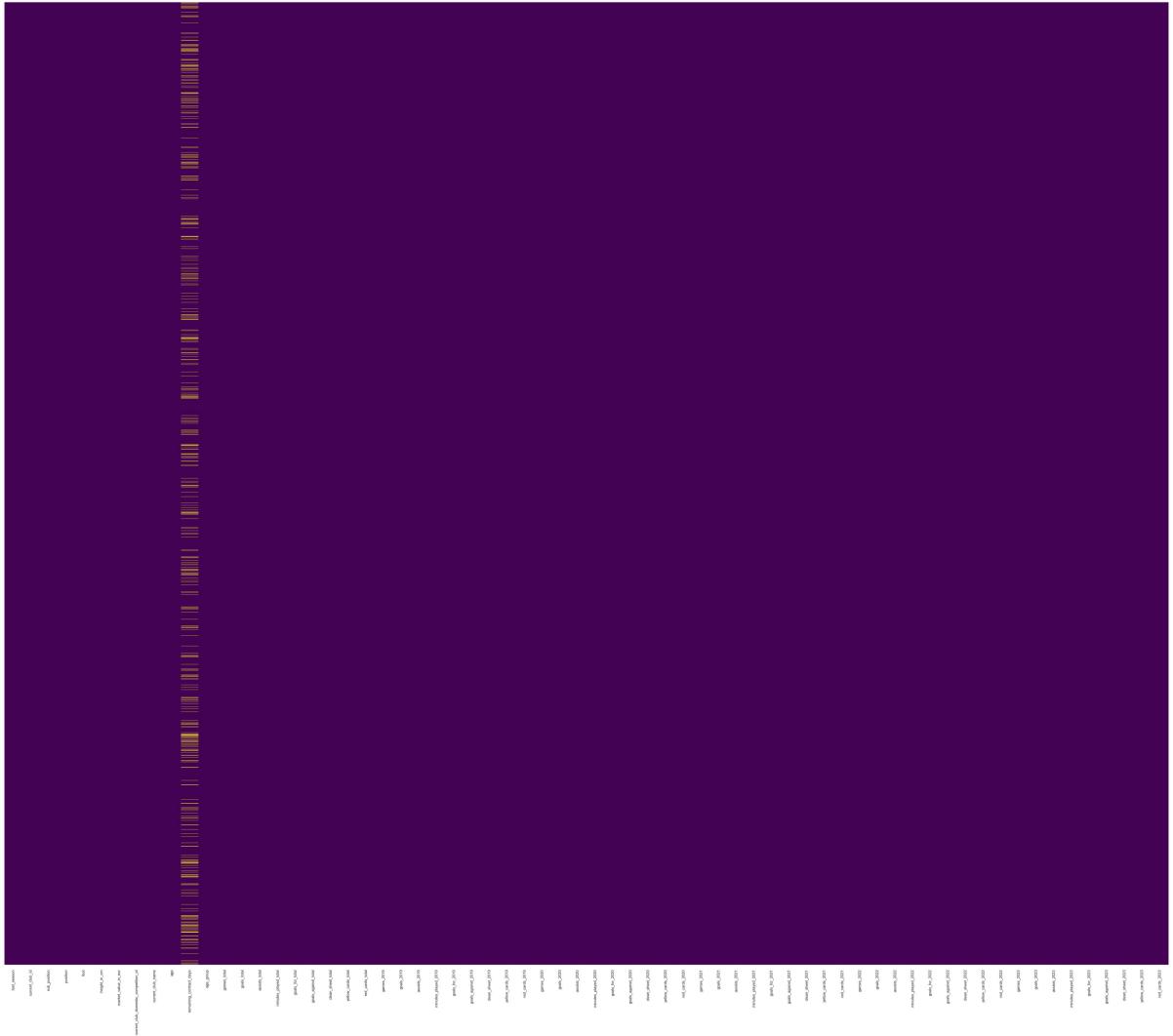
        else:
            return 'right'

    else:
        return foot
```

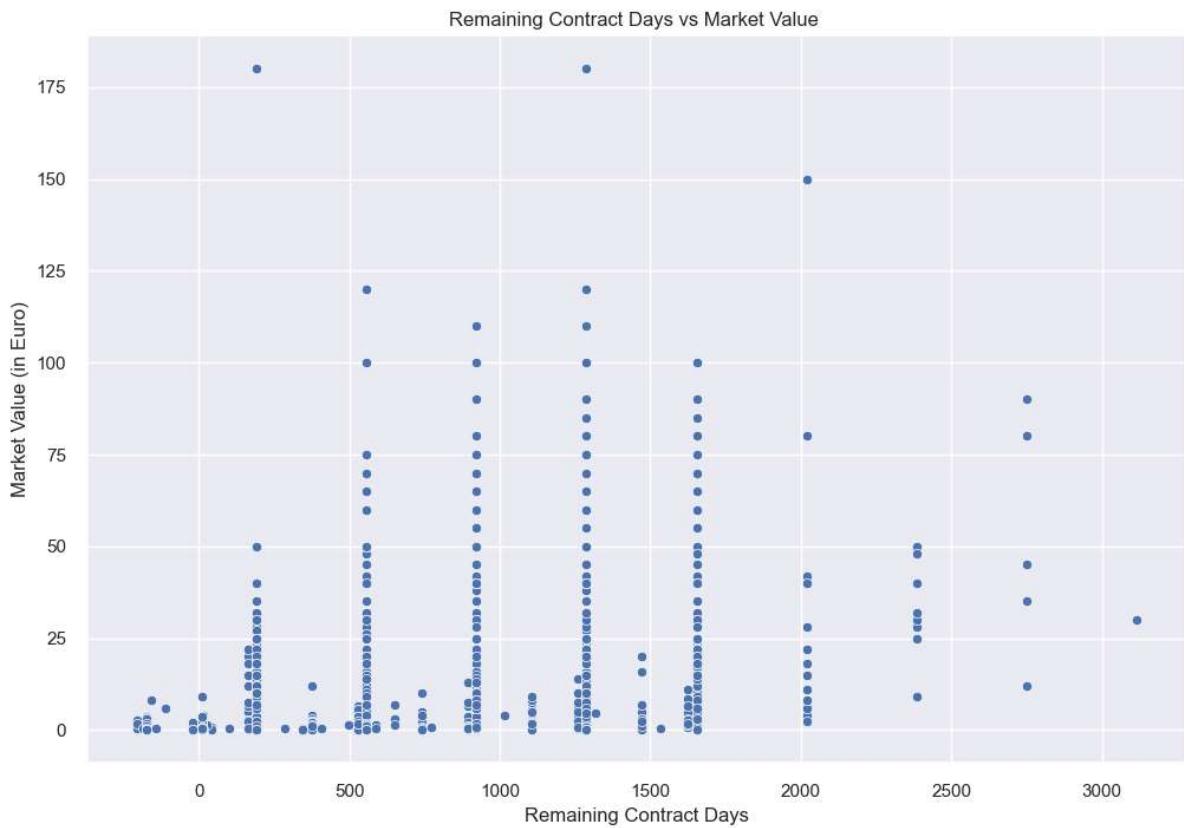
In [151]: # Applying the impute\_height function to the dataframe  
df4['foot'] = df4[['foot', 'position', 'sub\_position']].apply(impute\_foot, axis=1)

In [152]: sns.heatmap(df4.isnull(), yticklabels=False, cbar=False, cmap='viridis')

Out[152]: <Axes: >



```
In [153...]:  
plt.figure(figsize=(12, 8))  
sns.scatterplot(x='remaining_contract_days', y= df3['market_value_in_eur']/10000000,  
plt.title('Remaining Contract Days vs Market Value ')  
plt.xlabel('Remaining Contract Days')  
plt.ylabel('Market Value (in Euro)')  
plt.show()
```



```
In [154... correlation = df4['remaining_contract_days'].corr(df3['market_value_in_eur'])
print(f"Correlation between remaining_contract_days and market_value_in_eur: {corre
Correlation between remaining_contract_days and market_value_in_eur: 0.34085381505
098444

In [155...
# Calculating the mean market values based on position and age category
mean_remaining_contract_days_values = df4.groupby(['market_value_in_eur'])['remainin
# Displaying the mean values
print(mean_remaining_contract_days_values)
```

|    | market_value_in_eur | remaining_contract_days |
|----|---------------------|-------------------------|
| 0  | 10000.0             | 558.000000              |
| 1  | 25000.0             | 463.381818              |
| 2  | 50000.0             | 464.503846              |
| 3  | 75000.0             | 478.814815              |
| 4  | 100000.0            | 442.108359              |
| 5  | 125000.0            | 356.000000              |
| 6  | 150000.0            | 429.861423              |
| 7  | 175000.0            | 429.884615              |
| 8  | 176886.8            | NaN                     |
| 9  | 180000.0            | 528.000000              |
| 10 | 186000.0            | NaN                     |
| 11 | 200000.0            | 418.831731              |
| 12 | 225000.0            | 436.414634              |
| 13 | 250000.0            | 424.431034              |
| 14 | 275000.0            | 399.583333              |
| 15 | 300000.0            | 417.438228              |
| 16 | 325000.0            | 531.230769              |
| 17 | 346632.7            | 365.906250              |
| 18 | 350000.0            | 427.793269              |
| 19 | 375000.0            | 451.750000              |
| 20 | 400000.0            | 481.479911              |
| 21 | 425000.0            | 581.000000              |
| 22 | 443725.5            | -173.000000             |
| 23 | 450000.0            | 493.759259              |
| 24 | 475000.0            | 309.666667              |
| 25 | 500000.0            | 521.254364              |
| 26 | 522411.7            | 558.000000              |
| 27 | 550000.0            | 444.179487              |
| 28 | 600000.0            | 476.108949              |
| 29 | 641563.3            | NaN                     |
| 30 | 650000.0            | 484.814815              |
| 31 | 700000.0            | 542.885593              |
| 32 | 715892.2            | NaN                     |
| 33 | 750000.0            | 586.320755              |
| 34 | 790010.7            | NaN                     |
| 35 | 800000.0            | 592.174089              |
| 36 | 850000.0            | 605.914286              |
| 37 | 900000.0            | 569.528000              |
| 38 | 950000.0            | 740.500000              |
| 39 | 1000000.0           | 594.446023              |
| 40 | 1093724.0           | 153.000000              |
| 41 | 1100000.0           | 485.000000              |
| 42 | 1200000.0           | 634.903743              |
| 43 | 1300000.0           | 481.609756              |
| 44 | 1400000.0           | 572.560000              |
| 45 | 1500000.0           | 583.622699              |
| 46 | 1579725.0           | 423.714286              |
| 47 | 1600000.0           | 472.117647              |
| 48 | 1700000.0           | 512.375000              |
| 49 | 1800000.0           | 571.346535              |
| 50 | 1836525.0           | NaN                     |
| 51 | 1900000.0           | 923.000000              |
| 52 | 2000000.0           | 624.398230              |
| 53 | 2100000.0           | 679.666667              |
| 54 | 2109582.0           | NaN                     |
| 55 | 2200000.0           | 663.346154              |
| 56 | 2209577.0           | 401.689655              |
| 57 | 2300000.0           | 740.571429              |
| 58 | 2400000.0           | 983.833333              |
| 59 | 2500000.0           | 669.604938              |
| 60 | 2519384.0           | 192.500000              |
|    |                     | 375.500000              |
|    |                     | 923.000000              |

|     |            |             |
|-----|------------|-------------|
| 63  | 2700000.0  | 591.181818  |
| 64  | 2715097.0  | 488.261538  |
| 65  | 2800000.0  | 714.041667  |
| 66  | 2989714.0  | 108.307692  |
| 67  | 3000000.0  | 673.625468  |
| 68  | 3100000.0  | 558.000000  |
| 69  | 3168564.0  | 192.800000  |
| 70  | 3200000.0  | 1014.250000 |
| 71  | 3300000.0  | 801.333333  |
| 72  | 3400000.0  | 679.666667  |
| 73  | 3500000.0  | 702.798387  |
| 74  | 3700000.0  | 862.166667  |
| 75  | 3761589.0  | 134.900000  |
| 76  | 3800000.0  | 723.909091  |
| 77  | 4000000.0  | 801.484211  |
| 78  | 4100000.0  | 558.000000  |
| 79  | 4200000.0  | 996.200000  |
| 80  | 4300000.0  | 375.500000  |
| 81  | 4500000.0  | 782.446429  |
| 82  | 4800000.0  | 649.250000  |
| 83  | 5000000.0  | 776.507692  |
| 84  | 5500000.0  | 801.461538  |
| 85  | 6000000.0  | 789.408451  |
| 86  | 6500000.0  | 863.809524  |
| 87  | 7000000.0  | 740.955056  |
| 88  | 7500000.0  | 908.714286  |
| 89  | 8000000.0  | 857.253846  |
| 90  | 8500000.0  | 1029.600000 |
| 91  | 9000000.0  | 956.378788  |
| 92  | 9500000.0  | 558.000000  |
| 93  | 10000000.0 | 858.383459  |
| 94  | 11000000.0 | 1126.812500 |
| 95  | 12000000.0 | 935.107843  |
| 96  | 13000000.0 | 1004.548387 |
| 97  | 14000000.0 | 1001.243243 |
| 98  | 15000000.0 | 1050.390000 |
| 99  | 16000000.0 | 916.153846  |
| 100 | 17000000.0 | 895.076923  |
| 101 | 18000000.0 | 995.723077  |
| 102 | 19000000.0 | 1105.500000 |
| 103 | 20000000.0 | 1079.066667 |
| 104 | 21000000.0 | 193.000000  |
| 105 | 22000000.0 | 1115.558824 |
| 106 | 23000000.0 | 193.000000  |
| 107 | 24000000.0 | 1507.600000 |
| 108 | 25000000.0 | 1048.852459 |
| 109 | 26000000.0 | 558.000000  |
| 110 | 27000000.0 | 923.000000  |
| 111 | 28000000.0 | 1251.900000 |
| 112 | 30000000.0 | 1129.076923 |
| 113 | 32000000.0 | 1091.692308 |
| 114 | 35000000.0 | 1194.976744 |
| 115 | 38000000.0 | 1235.857143 |
| 116 | 40000000.0 | 1186.441860 |
| 117 | 42000000.0 | 1242.625000 |
| 118 | 45000000.0 | 1236.071429 |
| 119 | 48000000.0 | 1532.000000 |
| 120 | 50000000.0 | 1027.476190 |
| 121 | 55000000.0 | 1421.363636 |
| 122 | 60000000.0 | 1314.571429 |
| 123 | 65000000.0 | 1197.166667 |
| 124 | 70000000.0 | 1085.555556 |
|     |            | 1151.250000 |
|     |            | 1653.625000 |

```
127      85000000.0      1410.000000
128      90000000.0      1516.625000
129      100000000.0     1105.666667
130      110000000.0     1105.500000
131      120000000.0      923.000000
132      150000000.0     2019.000000
133      180000000.0      740.500000
```

```
In [156]: df4['remaining_contract_days'].describe()
```

```
Out[156]: count    8945.000000
mean     609.395081
std      459.677277
min     -203.000000
25%     193.000000
50%     558.000000
75%     923.000000
max     3115.000000
Name: remaining_contract_days, dtype: float64
```

```
In [157]: mean_remaining_days = df4['remaining_contract_days'].mean().astype(int)
```

```
In [158]: df5 = df4.copy()
```

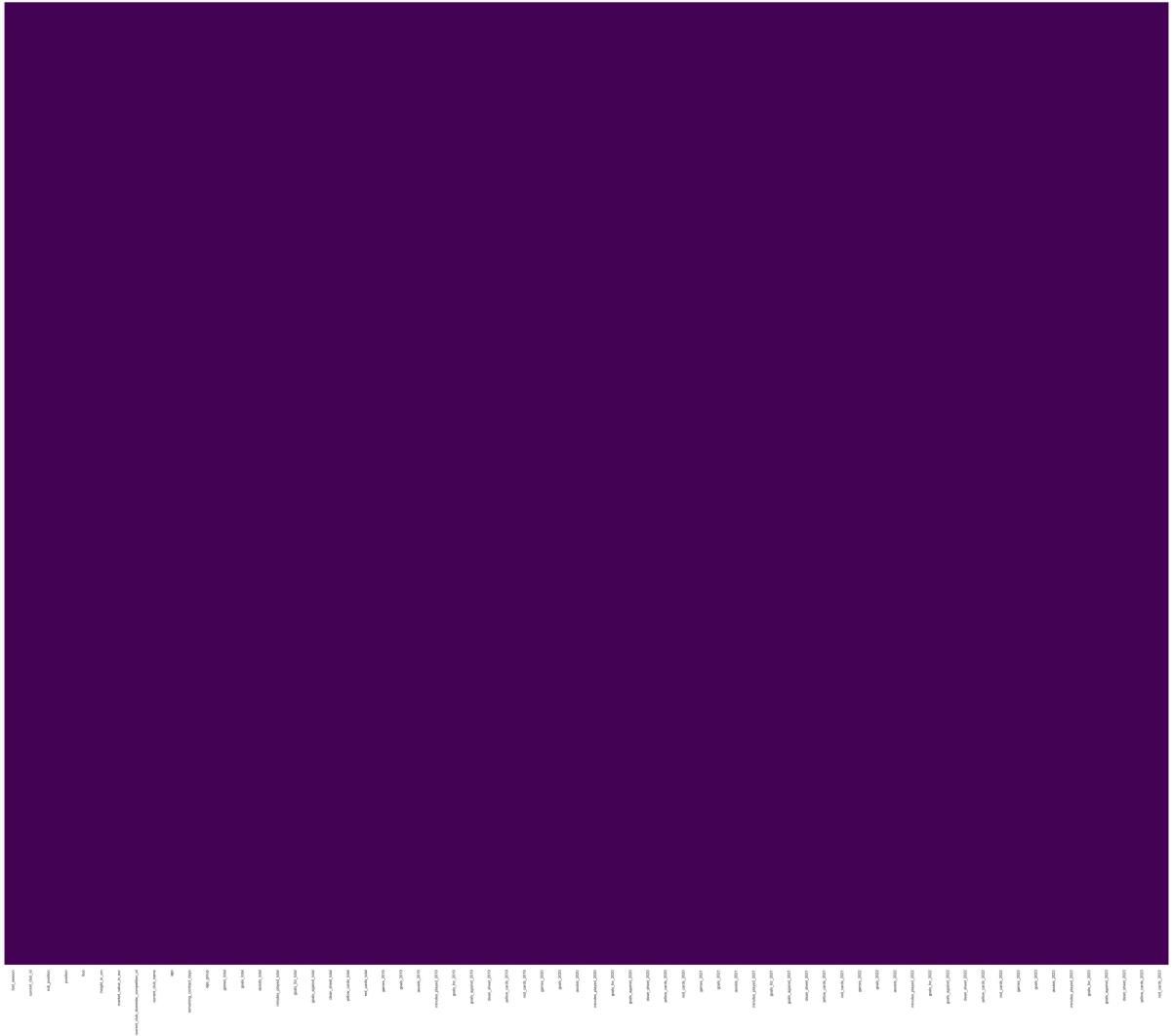
```
In [159]: df5['remaining_contract_days'].fillna(mean_remaining_days, inplace=True)
```

```
In [160]: df5.shape
```

```
Out[160]: (10169, 66)
```

```
In [161]: sns.heatmap(df5.isnull(), yticklabels=False, cbar=False, cmap='viridis')
```

```
Out[161]: <Axes: >
```



In [162...]

```
df5.isnull().sum()
```

```
Out[162]: last_season          0
           current_club_id      0
           sub_position         0
           position             0
           foot                 0
           height_in_cm          0
           market_value_in_eur    0
           current_club Domestic_competition_id 0
           current_club_name      0
           age                  0
           remaining_contract_days 0
           age_group            0
           games_total           0
           goals_total           0
           assists_total          0
           minutes_played_total   0
           goals_for_total        0
           goals_against_total     0
           clean_sheet_total       0
           yellow_cards_total      0
           red_cards_total          0
           games_2019              0
           goals_2019              0
           assists_2019             0
           minutes_played_2019      0
           goals_for_2019           0
           goals_against_2019        0
           clean_sheet_2019          0
           yellow_cards_2019         0
           red_cards_2019            0
           games_2020              0
           goals_2020              0
           assists_2020             0
           minutes_played_2020      0
           goals_for_2020           0
           goals_against_2020        0
           clean_sheet_2020          0
           yellow_cards_2020         0
           red_cards_2020            0
           games_2021              0
           goals_2021              0
           assists_2021             0
           minutes_played_2021      0
           goals_for_2021           0
           goals_against_2021        0
           clean_sheet_2021          0
           yellow_cards_2021         0
           red_cards_2021            0
           games_2022              0
           goals_2022              0
           assists_2022             0
           minutes_played_2022      0
           goals_for_2022           0
           goals_against_2022        0
           clean_sheet_2022          0
           yellow_cards_2022         0
           red_cards_2022            0
           games_2023              0
           goals_2023              0
           assists_2023             0
           minutes_played_2023      0
           goals_for_2023           0
                                         0
                                         0
```

```
yellow_cards_2023          0  
red_cards_2023             0  
dtype: int64
```

In [163]:

```
df5.nunique()
```

|           |                                      |      |
|-----------|--------------------------------------|------|
| Out[163]: | last_season                          | 2    |
|           | current_club_id                      | 274  |
|           | sub_position                         | 13   |
|           | position                             | 4    |
|           | foot                                 | 3    |
|           | height_in_cm                         | 59   |
|           | market_value_in_eur                  | 134  |
|           | current_club Domestic_competition_id | 14   |
|           | current_club_name                    | 274  |
|           | age                                  | 28   |
|           | remaining_contract_days              | 49   |
|           | age_group                            | 5    |
|           | games_total                          | 205  |
|           | goals_total                          | 91   |
|           | assists_total                        | 63   |
|           | minutes_played_total                 | 5227 |
|           | goals_for_total                      | 410  |
|           | goals_against_total                  | 256  |
|           | clean_sheet_total                    | 92   |
|           | yellow_cards_total                   | 55   |
|           | red_cards_total                      | 6    |
|           | games_2019                           | 57   |
|           | goals_2019                           | 35   |
|           | assists_2019                         | 24   |
|           | minutes_played_2019                  | 2203 |
|           | goals_for_2019                       | 128  |
|           | goals_against_2019                   | 83   |
|           | clean_sheet_2019                     | 27   |
|           | yellow_cards_2019                    | 21   |
|           | red_cards_2019                       | 3    |
|           | games_2020                           | 56   |
|           | goals_2020                           | 38   |
|           | assists_2020                         | 22   |
|           | minutes_played_2020                  | 2372 |
|           | goals_for_2020                       | 125  |
|           | goals_against_2020                   | 86   |
|           | clean_sheet_2020                     | 31   |
|           | yellow_cards_2020                    | 18   |
|           | red_cards_2020                       | 3    |
|           | games_2021                           | 55   |
|           | goals_2021                           | 35   |
|           | assists_2021                         | 23   |
|           | minutes_played_2021                  | 2610 |
|           | goals_for_2021                       | 130  |
|           | goals_against_2021                   | 84   |
|           | clean_sheet_2021                     | 27   |
|           | yellow_cards_2021                    | 19   |
|           | red_cards_2021                       | 3    |
|           | games_2022                           | 58   |
|           | goals_2022                           | 33   |
|           | assists_2022                         | 22   |
|           | minutes_played_2022                  | 2839 |
|           | goals_for_2022                       | 126  |
|           | goals_against_2022                   | 82   |
|           | clean_sheet_2022                     | 29   |
|           | yellow_cards_2022                    | 20   |
|           | red_cards_2022                       | 4    |
|           | games_2023                           | 21   |
|           | goals_2023                           | 15   |
|           | assists_2023                         | 10   |
|           | minutes_played_2023                  | 1118 |
|           | goals_for_2023                       | 51   |
|           | clean_sheet_2023                     | 36   |

```
yellow_cards_2023  
red_cards_2023  
dtype: int64
```

```
8  
3
```

In [164]: [df5.info\(\)](#)

```

<class 'pandas.core.frame.DataFrame'>
Index: 10169 entries, 72 to 30129
Data columns (total 66 columns):
 #   Column           Non-Null Count Dtype  
 --- 
 0   last_season      10169 non-null  int64   
 1   current_club_id  10169 non-null  int64   
 2   sub_position     10169 non-null  object  
 3   position         10169 non-null  object  
 4   foot              10169 non-null  object  
 5   height_in_cm    10169 non-null  float64 
 6   market_value_in_eur 10169 non-null  float64 
 7   current_club_domestic_competition_id 10169 non-null  object  
 8   current_club_name 10169 non-null  object  
 9   age               10169 non-null  int64   
 10  remaining_contract_days 10169 non-null  float64 
 11  age_group        10169 non-null  object  
 12  games_total      10169 non-null  int64   
 13  goals_total      10169 non-null  int64   
 14  assists_total    10169 non-null  int64   
 15  minutes_played_total 10169 non-null  int64   
 16  goals_for_total  10169 non-null  int64   
 17  goals_against_total 10169 non-null  int64   
 18  clean_sheet_total 10169 non-null  int64   
 19  yellow_cards_total 10169 non-null  int64   
 20  red_cards_total  10169 non-null  int64   
 21  games_2019       10169 non-null  float64 
 22  goals_2019       10169 non-null  float64 
 23  assists_2019     10169 non-null  float64 
 24  minutes_played_2019 10169 non-null  float64 
 25  goals_for_2019   10169 non-null  float64 
 26  goals_against_2019 10169 non-null  float64 
 27  clean_sheet_2019 10169 non-null  float64 
 28  yellow_cards_2019 10169 non-null  float64 
 29  red_cards_2019   10169 non-null  float64 
 30  games_2020       10169 non-null  float64 
 31  goals_2020       10169 non-null  float64 
 32  assists_2020     10169 non-null  float64 
 33  minutes_played_2020 10169 non-null  float64 
 34  goals_for_2020   10169 non-null  float64 
 35  goals_against_2020 10169 non-null  float64 
 36  clean_sheet_2020 10169 non-null  float64 
 37  yellow_cards_2020 10169 non-null  float64 
 38  red_cards_2020   10169 non-null  float64 
 39  games_2021       10169 non-null  float64 
 40  goals_2021       10169 non-null  float64 
 41  assists_2021     10169 non-null  float64 
 42  minutes_played_2021 10169 non-null  float64 
 43  goals_for_2021   10169 non-null  float64 
 44  goals_against_2021 10169 non-null  float64 
 45  clean_sheet_2021 10169 non-null  float64 
 46  yellow_cards_2021 10169 non-null  float64 
 47  red_cards_2021   10169 non-null  float64 
 48  games_2022       10169 non-null  float64 
 49  goals_2022       10169 non-null  int64   
 50  assists_2022     10169 non-null  int64   
 51  minutes_played_2022 10169 non-null  float64 
 52  goals_for_2022   10169 non-null  float64 
 53  goals_against_2022 10169 non-null  float64 
 54  clean_sheet_2022 10169 non-null  int64   
 55  yellow_cards_2022 10169 non-null  int64   
 56  red_cards_2022   10169 non-null  int64

```

```
59 assists_2023                      10169 non-null float64
60 minutes_played_2023                 10169 non-null float64
61 goals_for_2023                     10169 non-null float64
62 goals_against_2023                  10169 non-null float64
63 clean_sheet_2023                   10169 non-null float64
64 yellow_cards_2023                  10169 non-null float64
65 red_cards_2023                     10169 non-null float64
dtypes: float64(43), int64(17), object(6)
memory usage: 5.2+ MB
```

## Casting float datatypes to integer

In [165...]

```
# Simple function to iterate through the dataset and cast float type to int
def float_to_int(df):
    for column in df.columns:
        if df[column].dtype == 'float64' and column not in ['height_in_cm', 'market'
            df[column] = df[column].astype(int)
    return df
```

In [166...]

```
# Applying the 'float_to_int' function on the dataset and storing the results to df
df6 = float_to_int(df5)
```

In [167...]

```
df6.info()
```

```

<class 'pandas.core.frame.DataFrame'>
Index: 10169 entries, 72 to 30129
Data columns (total 66 columns):
 #   Column           Non-Null Count Dtype  
 --- 
 0   last_season      10169 non-null  int64   
 1   current_club_id  10169 non-null  int64   
 2   sub_position     10169 non-null  object  
 3   position         10169 non-null  object  
 4   foot              10169 non-null  object  
 5   height_in_cm    10169 non-null  float64 
 6   market_value_in_eur 10169 non-null  float64 
 7   current_club_domestic_competition_id 10169 non-null  object  
 8   current_club_name 10169 non-null  object  
 9   age               10169 non-null  int64   
 10  remaining_contract_days 10169 non-null  int32   
 11  age_group        10169 non-null  object  
 12  games_total      10169 non-null  int64   
 13  goals_total      10169 non-null  int64   
 14  assists_total    10169 non-null  int64   
 15  minutes_played_total 10169 non-null  int64   
 16  goals_for_total  10169 non-null  int64   
 17  goals_against_total 10169 non-null  int64   
 18  clean_sheet_total 10169 non-null  int64   
 19  yellow_cards_total 10169 non-null  int64   
 20  red_cards_total  10169 non-null  int64   
 21  games_2019       10169 non-null  int32   
 22  goals_2019       10169 non-null  int32   
 23  assists_2019     10169 non-null  int32   
 24  minutes_played_2019 10169 non-null  int32   
 25  goals_for_2019   10169 non-null  int32   
 26  goals_against_2019 10169 non-null  int32   
 27  clean_sheet_2019 10169 non-null  int32   
 28  yellow_cards_2019 10169 non-null  int32   
 29  red_cards_2019   10169 non-null  int32   
 30  games_2020       10169 non-null  int32   
 31  goals_2020       10169 non-null  int32   
 32  assists_2020     10169 non-null  int32   
 33  minutes_played_2020 10169 non-null  int32   
 34  goals_for_2020   10169 non-null  int32   
 35  goals_against_2020 10169 non-null  int32   
 36  clean_sheet_2020 10169 non-null  int32   
 37  yellow_cards_2020 10169 non-null  int32   
 38  red_cards_2020   10169 non-null  int32   
 39  games_2021       10169 non-null  int32   
 40  goals_2021       10169 non-null  int32   
 41  assists_2021     10169 non-null  int32   
 42  minutes_played_2021 10169 non-null  int32   
 43  goals_for_2021   10169 non-null  int32   
 44  goals_against_2021 10169 non-null  int32   
 45  clean_sheet_2021 10169 non-null  int32   
 46  yellow_cards_2021 10169 non-null  int32   
 47  red_cards_2021   10169 non-null  int32   
 48  games_2022       10169 non-null  int32   
 49  goals_2022       10169 non-null  int64   
 50  assists_2022     10169 non-null  int64   
 51  minutes_played_2022 10169 non-null  int32   
 52  goals_for_2022   10169 non-null  int32   
 53  goals_against_2022 10169 non-null  int32   
 54  clean_sheet_2022 10169 non-null  int64   
 55  yellow_cards_2022 10169 non-null  int64   
 56  red_cards_2022   10169 non-null  int64

```

```
59 assists_2023          10169 non-null int32
60 minutes_played_2023    10169 non-null int32
61 goals_for_2023         10169 non-null int32
62 goals_against_2023     10169 non-null int32
63 clean_sheet_2023       10169 non-null int32
64 yellow_cards_2023      10169 non-null int32
65 red_cards_2023         10169 non-null int32
dtypes: float64(2), int32(41), int64(17), object(6)
memory usage: 3.6+ MB
```

## Removing Duplicate Entries

```
In [168]: # Searching for duplicate rows and saving them to duplicateRows object
duplicateRows = df6[df6.duplicated()]
```

```
In [169]: # Viewing the duplicated entries
duplicateRows.head()
```

```
Out[169]:   last_season current_club_id sub_position position foot height_in_cm market_value_in_eu
              19948           2022            399        Central Midfield right      179.848057             2715097.
```



```
In [170]: #Checking the shape of df6 before dropping duplicates
df6.shape
```

```
Out[170]: (10169, 66)
```

```
In [171]: # Dropping duplicates and saving the new dataset as df6
df6 = df6.drop_duplicates()
```

```
In [172]: # Confirming the new shape after dropping the duplicated row
df6.shape
```

```
Out[172]: (10168, 66)
```

## More Feature Engineering

### Feature Generation

Creating a feature for minutes per a goal for each of the last five seasons

```
In [173]: # Creating a new copy of the dataframe incase we need to refer back to df6
df7 = df6.copy()
```

```
In [174]: df7['mins_per_goal_2023'] = df7['minutes_played_2023']/df7['goals_2023']
```

```
In [175]: df7['mins_per_goal_2022'] = df7['minutes_played_2022']/df7['goals_2022']
```

```
In [176]: df7['mins_per_goal_2021'] = df7['minutes_played_2021']/df7['goals_2021']
```

```
In [177]: df7['mins_per_goal_2020'] = df7['minutes_played_2020']/df7['goals_2020']
```

```
In [178]: df7['mins_per_goal_2019'] = df7['minutes_played_2019']/df7['goals_2019']
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

Creating a feature for minutes per a goal for total goals and minutes over the last five seasons.

```
In [179... df7['mins_per_goal_total'] = df7['minutes_played_total']/df7['goals_total']
```

Creating a feature for minutes per an assist for each of the last five seasons

```
In [180... df7['mins_per_assist_2023'] = df7['minutes_played_2023']/df7['assists_2023']
```

```
In [181... df7['mins_per_assist_2022'] = df7['minutes_played_2022']/df7['assists_2022']
```

```
In [182... df7['mins_per_assist_2021'] = df7['minutes_played_2021']/df7['assists_2021']
```

```
In [183... df7['mins_per_assist_2020'] = df7['minutes_played_2020']/df7['assists_2020']
```

```
In [184... df7['mins_per_assist_2019'] = df7['minutes_played_2019']/df7['assists_2019']
```

Creating a feature for minutes per an assist for total assists and minutes over the last five seasons.

```
In [185... df7['mins_per_assist_total'] = df7['minutes_played_total']/df7['assists_total']
```

```
In [186... df7.isnull().sum()
```

```
Out[186]: last_season          0
           current_club_id      0
           sub_position         0
           position             0
           foot                 0
           height_in_cm          0
           market_value_in_eur    0
           current_club Domestic_competition_id 0
           current_club_name      0
           age                  0
           remaining_contract_days 0
           age_group            0
           games_total           0
           goals_total           0
           assists_total          0
           minutes_played_total   0
           goals_for_total        0
           goals_against_total     0
           clean_sheet_total       0
           yellow_cards_total      0
           red_cards_total          0
           games_2019              0
           goals_2019              0
           assists_2019             0
           minutes_played_2019      0
           goals_for_2019           0
           goals_against_2019        0
           clean_sheet_2019          0
           yellow_cards_2019         0
           red_cards_2019            0
           games_2020              0
           goals_2020              0
           assists_2020             0
           minutes_played_2020      0
           goals_for_2020           0
           goals_against_2020        0
           clean_sheet_2020          0
           yellow_cards_2020         0
           red_cards_2020            0
           games_2021              0
           goals_2021              0
           assists_2021             0
           minutes_played_2021      0
           goals_for_2021           0
           goals_against_2021        0
           clean_sheet_2021          0
           yellow_cards_2021         0
           red_cards_2021            0
           games_2022              0
           goals_2022              0
           assists_2022             0
           minutes_played_2022      0
           goals_for_2022           0
           goals_against_2022        0
           clean_sheet_2022          0
           yellow_cards_2022         0
           red_cards_2022            0
           games_2023              0
           goals_2023              0
           assists_2023             0
           minutes_played_2023      0
           goals_for_2023           0
                                         0
                                         0
```

```
yellow_cards_2023          0
red_cards_2023              0
mins_per_goal_2023         4555
mins_per_goal_2022         3118
mins_per_goal_2021         5158
mins_per_goal_2020         5810
mins_per_goal_2019         6434
mins_per_goal_total        1335
mins_per_assist_2023       4555
mins_per_assist_2022       3118
mins_per_assist_2021       5158
mins_per_assist_2020       5810
mins_per_assist_2019       6434
mins_per_assist_total      1335
dtype: int64
```

In [187]: df7.head(50)

|     | last_season | current_club_id | sub_position       | position   | foot  | height_in_cm | market_value_in_eu |
|-----|-------------|-----------------|--------------------|------------|-------|--------------|--------------------|
| 72  | 2022        | 418             | Centre-Forward     | Attack     | right | 185.0        | 2500000.           |
| 88  | 2023        | 678             | Goalkeeper         | Goalkeeper | right | 188.0        | 200000.            |
| 117 | 2023        | 367             | Attacking Midfield | Midfield   | right | 180.0        | 2500000.           |
| 132 | 2022        | 1095            | Goalkeeper         | Goalkeeper | right | 184.0        | 200000.            |
| 135 | 2022        | 1519            | Centre-Forward     | Attack     | left  | 185.0        | 200000.            |
| 152 | 2023        | 2425            | Defensive Midfield | Midfield   | right | 182.0        | 100000.            |
| 161 | 2023        | 371             | Goalkeeper         | Goalkeeper | right | 196.0        | 1500000.           |
| 163 | 2022        | 940             | Goalkeeper         | Goalkeeper | left  | 187.0        | 300000.            |
| 175 | 2023        | 58              | Centre-Back        | Defender   | left  | 189.0        | 1500000.           |
| 178 | 2023        | 511             | Goalkeeper         | Goalkeeper | right | 190.0        | 150000.            |
| 181 | 2023        | 336             | Goalkeeper         | Goalkeeper | left  | 190.0        | 1000000.           |
| 186 | 2023        | 380             | Centre-Back        | Defender   | right | 195.0        | 400000.            |
| 187 | 2023        | 294             | Right Winger       | Attack     | left  | 180.0        | 8000000.           |
| 190 | 2022        | 2239            | Centre-Forward     | Attack     | right | 187.0        | 350000.            |
| 191 | 2022        | 703             | Left Winger        | Attack     | left  | 175.0        | 1800000.           |
| 200 | 2023        | 903             | Centre-Forward     | Attack     | right | 180.0        | 400000.            |
| 217 | 2022        | 931             | Centre-Forward     | Attack     | right | 189.0        | 28000000.          |
| 228 | 2023        | 105             | Left-Back          | Defender   | left  | 172.0        | 600000.            |
| 232 | 2023        | 2919            | Centre-Back        | Defender   | right | 180.0        | 1500000.           |
| 238 | 2023        | 347             | Goalkeeper         | Goalkeeper | right | 184.0        | 500000.            |
| 240 | 2022        | 533             | Defensive Midfield | Midfield   | right | 180.0        | 1000000.           |
| 246 | 2023        | 48332           | Central Midfield   | Midfield   | right | 182.0        | 150000.            |
|     |             |                 | Back               | Defender   | left  | 180.0        | 40000000.          |

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

| last_season      | current_club_id | sub_position | position           | foot       | height_in_cm | market_value_in_eu |           |
|------------------|-----------------|--------------|--------------------|------------|--------------|--------------------|-----------|
| 255              | 2023            | 234          | Centre-Back        | Defender   | right        | 183.0              | 5000000.  |
| 258              | 2023            | 232          | Right Midfield     | Midfield   | right        | 177.0              | 2000000.  |
| 261              | 2023            | 3385         | Centre-Back        | Defender   | right        | 190.0              | 100000.   |
| 262              | 2022            | 667          | Centre-Back        | Defender   | left         | 185.0              | 522411.   |
| 277              | 2022            | 1025         | Attacking Midfield | Midfield   | right        | 182.0              | 2500000.  |
| 281              | 2023            | 12321        | Centre-Back        | Defender   | right        | 192.0              | 1500000.  |
| 283              | 2023            | 398          | Right Winger       | Attack     | both         | 167.0              | 1500000.  |
| 287              | 2022            | 2687         | Centre-Back        | Defender   | right        | 186.0              | 600000.   |
| 294              | 2023            | 1894         | Attacking Midfield | Midfield   | right        | 172.0              | 150000.   |
| 295              | 2023            | 8024         | Central Midfield   | Midfield   | left         | 183.0              | 250000.   |
| 302              | 2022            | 678          | Left-Back          | Defender   | both         | 188.0              | 522411.   |
| 313              | 2023            | 273          | Defensive Midfield | Midfield   | left         | 194.0              | 4000000.  |
| 319              | 2022            | 667          | Goalkeeper         | Goalkeeper | right        | 185.0              | 150000.   |
| 323              | 2023            | 3948         | Goalkeeper         | Goalkeeper | left         | 186.0              | 1500000.  |
| 326              | 2023            | 2578         | Central Midfield   | Midfield   | right        | 179.0              | 350000.   |
| 329              | 2023            | 1124         | Right Winger       | Attack     | right        | 175.0              | 350000.   |
| 332              | 2022            | 33           | Centre-Back        | Defender   | right        | 189.0              | 900000.   |
| 336              | 2023            | 873          | Right-Back         | Defender   | right        | 175.0              | 2000000.  |
| 339              | 2023            | 1083         | Attacking Midfield | Midfield   | right        | 177.0              | 800000.   |
| 340              | 2023            | 418          | Centre-Back        | Defender   | right        | 190.0              | 32000000. |
| 344              | 2022            | 683          | Attacking Midfield | Midfield   | left         | 181.0              | 6000000.  |
| 349              | 2023            | 1245         | Centre-Forward     | Attack     | right        | 184.0              | 350000.   |
| Central Midfield |                 |              |                    | Midfield   | left         | 175.0              | 400000.   |

| last_season | current_club_id | sub_position | position           | foot           | height_in_cm | market_value_in_eu |
|-------------|-----------------|--------------|--------------------|----------------|--------------|--------------------|
| 359         | 2023            | 58           | Defensive Midfield | Midfield right | 178.0        | 5000000.           |
| 370         | 2023            | 60551        | Right Midfield     | Midfield both  | 166.0        | 600000.            |
| 372         | 2022            | 1096         | Left Winger        | Attack right   | 181.0        | 2519384.           |
| 383         | 2022            | 3205         | Central            | Midfield left  | 179.0        | 1300000.           |

The new features contain Nan and inf values as some plays have not scored or assisted or played any minutes. 0/0 = Nan, 0/number = inf. Our ML models cannot handle these values. We will impute them with a large value for infinity.

In [188...]

```
# Setting a Large number in scientific notation for infinity
#INF = 1e400

nan_inf_features_list = ['mins_per_goal_2023', 'mins_per_goal_2022', 'mins_per_goal_2021']

# Replacing the NaN and inf values with 1e400
for feature in nan_inf_features_list:
    df7[feature].replace([np.nan, np.inf], 1e400, inplace=True)
```

In [189...]

```
df7.isnull().sum()
```

```
Out[189]: last_season          0
           current_club_id      0
           sub_position         0
           position             0
           foot                 0
           height_in_cm          0
           market_value_in_eur    0
           current_club Domestic_competition_id 0
           current_club_name      0
           age                  0
           remaining_contract_days 0
           age_group            0
           games_total           0
           goals_total           0
           assists_total          0
           minutes_played_total   0
           goals_for_total        0
           goals_against_total     0
           clean_sheet_total       0
           yellow_cards_total      0
           red_cards_total          0
           games_2019              0
           goals_2019              0
           assists_2019             0
           minutes_played_2019      0
           goals_for_2019           0
           goals_against_2019        0
           clean_sheet_2019          0
           yellow_cards_2019         0
           red_cards_2019            0
           games_2020              0
           goals_2020              0
           assists_2020             0
           minutes_played_2020      0
           goals_for_2020           0
           goals_against_2020        0
           clean_sheet_2020          0
           yellow_cards_2020         0
           red_cards_2020            0
           games_2021              0
           goals_2021              0
           assists_2021             0
           minutes_played_2021      0
           goals_for_2021           0
           goals_against_2021        0
           clean_sheet_2021          0
           yellow_cards_2021         0
           red_cards_2021            0
           games_2022              0
           goals_2022              0
           assists_2022             0
           minutes_played_2022      0
           goals_for_2022           0
           goals_against_2022        0
           clean_sheet_2022          0
           yellow_cards_2022         0
           red_cards_2022            0
           games_2023              0
           goals_2023              0
           assists_2023             0
           minutes_played_2023      0
           goals_for_2023           0
                                         0
                                         0
```

```
yellow_cards_2023          0
red_cards_2023              0
mins_per_goal_2023          0
mins_per_goal_2022          0
mins_per_goal_2021          0
mins_per_goal_2020          0
mins_per_goal_2019          0
mins_per_goal_total         0
mins_per_assist_2023        0
mins_per_assist_2022        0
mins_per_assist_2021        0
mins_per_assist_2020        0
mins_per_assist_2019        0
mins_per_assist_total       0
dtype: int64
```

## Label Encoding

### One-hot Label Encoding (dummies)

One-hot Label Encoding will be used instead of factorising variables or standard label encoding as there is no order or rank between the categorical variables. It will create a new binary feature for each category and assigns a value of 1 or 0 depending on whether that player belongs to that category.

In [223...]

```
df7.info()
```

```

<class 'pandas.core.frame.DataFrame'>
Index: 10168 entries, 72 to 30129
Data columns (total 78 columns):
 #   Column           Non-Null Count Dtype  
 --- 
 0   last_season      10168 non-null  int64   
 1   current_club_id  10168 non-null  int64   
 2   sub_position     10168 non-null  object  
 3   position         10168 non-null  object  
 4   foot              10168 non-null  object  
 5   height_in_cm    10168 non-null  float64 
 6   market_value_in_eur 10168 non-null  float64 
 7   current_club_domestic_competition_id 10168 non-null  object  
 8   current_club_name 10168 non-null  object  
 9   age               10168 non-null  int64   
 10  remaining_contract_days 10168 non-null  int32   
 11  age_group        10168 non-null  object  
 12  games_total      10168 non-null  int64   
 13  goals_total      10168 non-null  int64   
 14  assists_total    10168 non-null  int64   
 15  minutes_played_total 10168 non-null  int64   
 16  goals_for_total  10168 non-null  int64   
 17  goals_against_total 10168 non-null  int64   
 18  clean_sheet_total 10168 non-null  int64   
 19  yellow_cards_total 10168 non-null  int64   
 20  red_cards_total  10168 non-null  int64   
 21  games_2019       10168 non-null  int32   
 22  goals_2019       10168 non-null  int32   
 23  assists_2019     10168 non-null  int32   
 24  minutes_played_2019 10168 non-null  int32   
 25  goals_for_2019   10168 non-null  int32   
 26  goals_against_2019 10168 non-null  int32   
 27  clean_sheet_2019 10168 non-null  int32   
 28  yellow_cards_2019 10168 non-null  int32   
 29  red_cards_2019   10168 non-null  int32   
 30  games_2020       10168 non-null  int32   
 31  goals_2020       10168 non-null  int32   
 32  assists_2020     10168 non-null  int32   
 33  minutes_played_2020 10168 non-null  int32   
 34  goals_for_2020   10168 non-null  int32   
 35  goals_against_2020 10168 non-null  int32   
 36  clean_sheet_2020 10168 non-null  int32   
 37  yellow_cards_2020 10168 non-null  int32   
 38  red_cards_2020   10168 non-null  int32   
 39  games_2021       10168 non-null  int32   
 40  goals_2021       10168 non-null  int32   
 41  assists_2021     10168 non-null  int32   
 42  minutes_played_2021 10168 non-null  int32   
 43  goals_for_2021   10168 non-null  int32   
 44  goals_against_2021 10168 non-null  int32   
 45  clean_sheet_2021 10168 non-null  int32   
 46  yellow_cards_2021 10168 non-null  int32   
 47  red_cards_2021   10168 non-null  int32   
 48  games_2022       10168 non-null  int32   
 49  goals_2022       10168 non-null  int64   
 50  assists_2022     10168 non-null  int64   
 51  minutes_played_2022 10168 non-null  int32   
 52  goals_for_2022   10168 non-null  int32   
 53  goals_against_2022 10168 non-null  int32   
 54  clean_sheet_2022 10168 non-null  int64   
 55  yellow_cards_2022 10168 non-null  int64   
 56  red_cards_2022   10168 non-null  int64

```

```

59 assists_2023           10168 non-null  int32
60 minutes_played_2023    10168 non-null  int32
61 goals_for_2023         10168 non-null  int32
62 goals_against_2023     10168 non-null  int32
63 clean_sheet_2023       10168 non-null  int32
64 yellow_cards_2023      10168 non-null  int32
65 red_cards_2023          10168 non-null  int32
66 mins_per_goal_2023     10168 non-null  float64
67 mins_per_goal_2022      10168 non-null  float64
68 mins_per_goal_2021      10168 non-null  float64
69 mins_per_goal_2020      10168 non-null  float64
70 mins_per_goal_2019      10168 non-null  float64
71 mins_per_goal_total     10168 non-null  float64
72 mins_per_assist_2023    10168 non-null  float64
73 mins_per_assist_2022    10168 non-null  float64
74 mins_per_assist_2021    10168 non-null  float64
75 mins_per_assist_2020    10168 non-null  float64
76 mins_per_assist_2019    10168 non-null  float64
77 mins_per_assist_total   10168 non-null  float64
dtypes: float64(14), int32(41), int64(17), object(6)
memory usage: 4.5+ MB

```

Initially a dataframe of all categorical variables was created but applying one-hot label encoding to it resulted in boolean type features. I thus decided to try splitting them into essential and non-essential categoricals.

```
In [234... df7['last_season'] = df7['last_season'].astype(str)
```

```
In [235... # Isolating the categorical features for encoding
dummies_df_essential_cats = df7[['last_season', 'sub_position', 'position', 'foot',
dummies_df_non_essential_cats = df7[['current_club Domestic_competition_id', 'curre
```

```
In [236... dummies_df_essential_cats.info()
```

```

<class 'pandas.core.frame.DataFrame'>
Index: 10168 entries, 72 to 30129
Data columns (total 5 columns):
 #   Column      Non-Null Count  Dtype  
---  -- 
 0   last_season  10168 non-null  object 
 1   sub_position 10168 non-null  object 
 2   position     10168 non-null  object 
 3   foot         10168 non-null  object 
 4   age_group    10168 non-null  object 
dtypes: object(5)
memory usage: 476.6+ KB

```

```
In [237... dummies_df_essential_cats = pd.get_dummies(dummies_df_essential_cats, dtype=int)
```

```
In [238... dummies_df_essential_cats.info()
```

```

<class 'pandas.core.frame.DataFrame'>
Index: 10168 entries, 72 to 30129
Data columns (total 27 columns):
 #   Column           Non-Null Count Dtype
 ---  -----
 0   last_season_2022    10168 non-null  int32
 1   last_season_2023    10168 non-null  int32
 2   sub_position_Attacking Midfield 10168 non-null  int32
 3   sub_position_Central Midfield 10168 non-null  int32
 4   sub_position_Centre-Back    10168 non-null  int32
 5   sub_position_Centre-Forward 10168 non-null  int32
 6   sub_position_Defensive Midfield 10168 non-null  int32
 7   sub_position_Goalkeeper    10168 non-null  int32
 8   sub_position_Left Midfield 10168 non-null  int32
 9   sub_position_Left Winger   10168 non-null  int32
 10  sub_position_Left-Back    10168 non-null  int32
 11  sub_position_Right Midfield 10168 non-null  int32
 12  sub_position_Right Winger 10168 non-null  int32
 13  sub_position_Right-Back   10168 non-null  int32
 14  sub_position_Second Striker 10168 non-null  int32
 15  position_Attack        10168 non-null  int32
 16  position_Defender      10168 non-null  int32
 17  position_Goalkeeper    10168 non-null  int32
 18  position_Midfield      10168 non-null  int32
 19  foot_both              10168 non-null  int32
 20  foot_left              10168 non-null  int32
 21  foot_right             10168 non-null  int32
 22  age_group_Ages 22 - 26  10168 non-null  int32
 23  age_group_Ages 27 - 32  10168 non-null  int32
 24  age_group_Ages 33 - 38  10168 non-null  int32
 25  age_group_Ages 39 - 44  10168 non-null  int32
 26  age_group_Ages Under 22 10168 non-null  int32
dtypes: int32(27)
memory usage: 1.1 MB

```

In [239]: `dummies_df_essential_cats.head()`

Out[239]:

|            | last_season_2022 | last_season_2023 | sub_position_Attacking Midfield | sub_position_Central Midfield | sub_position |
|------------|------------------|------------------|---------------------------------|-------------------------------|--------------|
| <b>72</b>  | 1                | 0                | 0                               | 0                             | 0            |
| <b>88</b>  | 0                | 1                | 0                               | 0                             | 0            |
| <b>117</b> | 0                | 1                | 1                               | 0                             | 0            |
| <b>132</b> | 1                | 0                | 0                               | 0                             | 0            |
| <b>135</b> | 1                | 0                | 0                               | 0                             | 0            |

In [240]: `dummies_df_essential_cats.shape`

Out[240]: (10168, 27)

In [241]: `dummies_df_non_essential_cats = pd.get_dummies(dummies_df_non_essential_cats, dtype=`

In [242]: `dummies_df_non_essential_cats.info()`

```
<class 'pandas.core.frame.DataFrame'>
Index: 10168 entries, 72 to 30129
Columns: 288 entries, current_club Domestic_competition_id_BE1 to current_club_name_Umraniyespor
dtypes: int32(288)
memory usage: 11.2 MB
```

```
In [243...]: dummies_df_non_essential_cats.shape
```

```
Out[243]: (10168, 288)
```

```
In [244...]: all_dummies = pd.concat([dummies_df_essential_cats, dummies_df_non_essential_cats])
```

```
In [245...]: all_dummies.shape
```

```
Out[245]: (10168, 315)
```

## Combining the dummies generated datasets to the original numerical features

Multiple datasets will be created and tested in our ML models.

A combined dataset of all categorical features following dummies...shape(10168, 386).

A dataset of numerical features and only 'foot', 'position', 'sub\_position' and 'age\_group' categories following one-hot label encoding...shape(10168, 98)

And another following feature selection...

```
In [246...]: numerical_features_df = df7.drop(['last_season', 'sub_position', 'position', 'foot'],
```

```
In [247...]: df8_dummies_essential = pd.concat([numerical_features_df, dummies_df_essential_cats])
```

```
In [248...]: df8_dummies_essential.info()
```

```

<class 'pandas.core.frame.DataFrame'>
Index: 10168 entries, 72 to 30129
Data columns (total 98 columns):
 #   Column           Non-Null Count Dtype
 ---  -- 
 0   current_club_id    10168 non-null int64
 1   height_in_cm       10168 non-null float64
 2   market_value_in_eur 10168 non-null float64
 3   age                10168 non-null int64
 4   remaining_contract_days 10168 non-null int32
 5   games_total        10168 non-null int64
 6   goals_total        10168 non-null int64
 7   assists_total      10168 non-null int64
 8   minutes_played_total 10168 non-null int64
 9   goals_for_total    10168 non-null int64
 10  goals_against_total 10168 non-null int64
 11  clean_sheet_total  10168 non-null int64
 12  yellow_cards_total 10168 non-null int64
 13  red_cards_total    10168 non-null int64
 14  games_2019         10168 non-null int32
 15  goals_2019         10168 non-null int32
 16  assists_2019       10168 non-null int32
 17  minutes_played_2019 10168 non-null int32
 18  goals_for_2019    10168 non-null int32
 19  goals_against_2019 10168 non-null int32
 20  clean_sheet_2019   10168 non-null int32
 21  yellow_cards_2019  10168 non-null int32
 22  red_cards_2019    10168 non-null int32
 23  games_2020         10168 non-null int32
 24  goals_2020         10168 non-null int32
 25  assists_2020       10168 non-null int32
 26  minutes_played_2020 10168 non-null int32
 27  goals_for_2020    10168 non-null int32
 28  goals_against_2020 10168 non-null int32
 29  clean_sheet_2020   10168 non-null int32
 30  yellow_cards_2020  10168 non-null int32
 31  red_cards_2020    10168 non-null int32
 32  games_2021         10168 non-null int32
 33  goals_2021         10168 non-null int32
 34  assists_2021       10168 non-null int32
 35  minutes_played_2021 10168 non-null int32
 36  goals_for_2021    10168 non-null int32
 37  goals_against_2021 10168 non-null int32
 38  clean_sheet_2021   10168 non-null int32
 39  yellow_cards_2021  10168 non-null int32
 40  red_cards_2021    10168 non-null int32
 41  games_2022         10168 non-null int32
 42  goals_2022         10168 non-null int64
 43  assists_2022       10168 non-null int64
 44  minutes_played_2022 10168 non-null int32
 45  goals_for_2022    10168 non-null int32
 46  goals_against_2022 10168 non-null int32
 47  clean_sheet_2022   10168 non-null int64
 48  yellow_cards_2022  10168 non-null int64
 49  red_cards_2022    10168 non-null int64
 50  games_2023         10168 non-null int32
 51  goals_2023         10168 non-null int32
 52  assists_2023       10168 non-null int32
 53  minutes_played_2023 10168 non-null int32
 54  goals_for_2023    10168 non-null int32
 55  goals_against_2023 10168 non-null int32
 56  clean_sheet_2023   10168 non-null int32

```

```

59 mins_per_goal_2023           10168 non-null float64
60 mins_per_goal_2022           10168 non-null float64
61 mins_per_goal_2021           10168 non-null float64
62 mins_per_goal_2020           10168 non-null float64
63 mins_per_goal_2019           10168 non-null float64
64 mins_per_goal_total          10168 non-null float64
65 mins_per_assist_2023         10168 non-null float64
66 mins_per_assist_2022         10168 non-null float64
67 mins_per_assist_2021         10168 non-null float64
68 mins_per_assist_2020         10168 non-null float64
69 mins_per_assist_2019         10168 non-null float64
70 mins_per_assist_total        10168 non-null float64
71 last_season_2022              10168 non-null int32
72 last_season_2023              10168 non-null int32
73 sub_position_Attacking Midfield 10168 non-null int32
74 sub_position_Central Midfield 10168 non-null int32
75 sub_position_Centre-Back      10168 non-null int32
76 sub_position_Centre-Forward   10168 non-null int32
77 sub_position_Defensive Midfield 10168 non-null int32
78 sub_position_Goalkeeper       10168 non-null int32
79 sub_position_Left Midfield   10168 non-null int32
80 sub_position_Left Winger     10168 non-null int32
81 sub_position_Left-Back        10168 non-null int32
82 sub_position_Right Midfield  10168 non-null int32
83 sub_position_Right Winger    10168 non-null int32
84 sub_position_Right-Back       10168 non-null int32
85 sub_position_Second Striker  10168 non-null int32
86 position_Attack              10168 non-null int32
87 position_Defender            10168 non-null int32
88 position_Goalkeeper          10168 non-null int32
89 position_Midfield            10168 non-null int32
90 foot_both                     10168 non-null int32
91 foot_left                     10168 non-null int32
92 foot_right                    10168 non-null int32
93 age_group_Ages 22 - 26        10168 non-null int32
94 age_group_Ages 27 - 32        10168 non-null int32
95 age_group_Ages 33 - 38        10168 non-null int32
96 age_group_Ages 39 - 44        10168 non-null int32
97 age_group_Ages Under 22      10168 non-null int32
dtypes: float64(14), int32(68), int64(16)
memory usage: 5.0 MB

```

In [249...]: df8\_dummies\_all = pd.concat([numerical\_features\_df, all\_dummies], axis=1)

In [250...]: df8\_dummies\_all.info()

```

<class 'pandas.core.frame.DataFrame'>
Index: 10168 entries, 72 to 30129
Columns: 386 entries, current_club_id to current_club_name_Ümraniyespor
dtypes: float64(14), int32(356), int64(16)
memory usage: 16.2 MB

```

## Correlation Between Features

### Pairplot between features and target variable

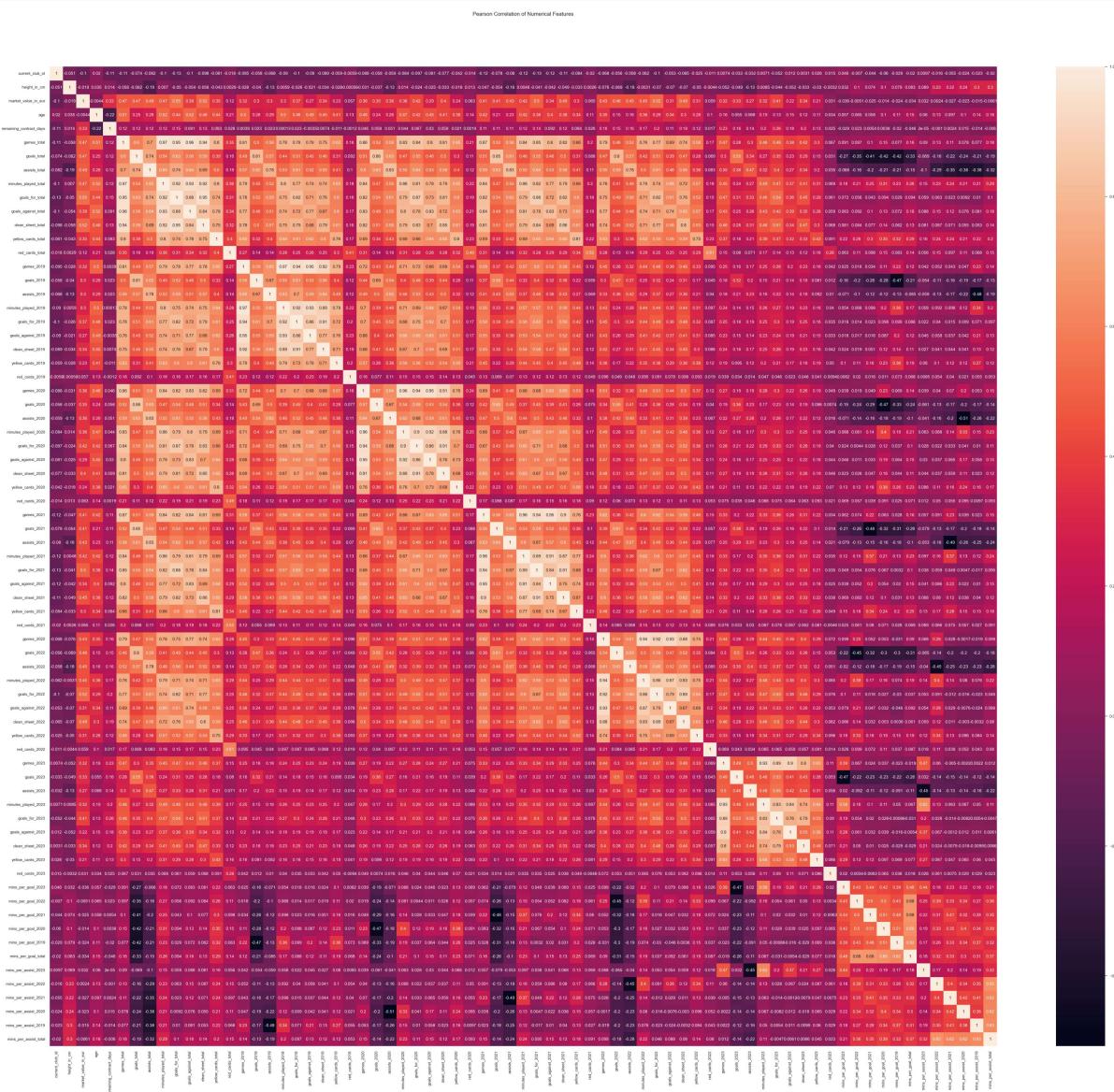
In [207...]: # sns.pairplot(df7, hue='market\_value\_in\_eur')
# plt.show()

### Pearson's Correlation Matrix of Numerical Features

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js

In [251...]

```
# Plot a heatmap of the correlation matrix of the numerical columns
sns.set(rc={"figure.figsize":(60, 50)})
sns.heatmap(numerical_features_df.corr(), annot = True)
plt.title('Pearson Correlation of Numerical Features', y=1.05, size=15)
cor = numerical_features_df.corr()
plt.show()
```



Correlation Matrix: The correlation coefficient has values between -1 to 1. A value closer to 0 implies weaker correlation (exact 0 implying no correlation) A value closer to 1 implies stronger positive correlation A value closer to -1 implies stronger negative correlation.

The Heatmap above is difficult to interpret due to its size and the number of features within the dataset.

In [252...]

```
# Correlation with target variable
cor_target = abs(cor['market_value_in_eur'])

# Selecting highly correlated features
relevant_features = cor_target[cor_target>0.4]
relevant_features
```

```
Out[252]: market_value_in_eur      1.000000
          games_total           0.470027
          goals_total            0.467667
          assists_total           0.486171
          minutes_played_total   0.471368
          goals_for_total         0.545848
          clean_sheet_total       0.519063
          goals_for_2020           0.422572
          games_2021              0.412321
          goals_2021               0.413660
          assists_2021              0.427364
          minutes_played_2021     0.416458
          goals_for_2021           0.495140
          clean_sheet_2021          0.447354
          games_2022              0.434351
          goals_2022               0.457946
          assists_2022              0.452212
          minutes_played_2022     0.446138
          goals_for_2022             0.518047
          clean_sheet_2022          0.488459
          goals_for_2023             0.406962
Name: market_value_in_eur, dtype: float64
```

### Pearson's Correlation between selected categorical features (foot, position, sub\_position, age\_group)

```
In [253... cor = df8_dummies_essential.corr()
```

```
In [254... # Correlation with target variable
cor_target = abs(cor['market_value_in_eur'])

# Selecting highly correlated features
relevant_features = cor_target[cor_target>0.3]
relevant_features
```

```
Out[254]: market_value_in_eur      1.000000
remaining_contract_days      0.334212
games_total                  0.470027
goals_total                  0.467667
assists_total                0.486171
minutes_played_total         0.471368
goals_for_total              0.545848
goals_against_total          0.392081
clean_sheet_total             0.519063
yellow_cards_total           0.352771
games_2019                   0.319025
assists_2019                 0.303601
minutes_played_2019          0.304588
goals_for_2019               0.365819
clean_sheet_2019              0.335958
games_2020                   0.361509
goals_2020                   0.349266
assists_2020                 0.359543
minutes_played_2020           0.357576
goals_for_2020               0.422572
clean_sheet_2020              0.395995
games_2021                   0.412321
goals_2021                   0.413660
assists_2021                 0.427364
minutes_played_2021           0.416458
goals_for_2021               0.495140
goals_against_2021            0.338094
clean_sheet_2021              0.447354
games_2022                   0.434351
goals_2022                   0.457946
assists_2022                 0.452212
minutes_played_2022           0.446138
goals_for_2022               0.518047
goals_against_2022            0.312198
clean_sheet_2022              0.488459
yellow_cards_2022             0.314594
games_2023                   0.319920
goals_2023                   0.329007
minutes_played_2023           0.324534
goals_for_2023               0.406962
clean_sheet_2023              0.338500
Name: market_value_in_eur, dtype: float64
```

Surprisingly, categories such as foot, position, sub-position and age-group do not seem to correlate highly with market-value.

## Pearson's Correlation between All Features

```
In [255...]: cor = df8_dummies_all.corr()
```

```
In [256...]: # Correlation with target variable
cor_target = abs(cor['market_value_in_eur'])

# Selecting highly correlated features
relevant_features = cor_target[cor_target > 0.1]
relevant_features
```

|           |  |          |
|-----------|--|----------|
| Out[256]: | current_club_id                          | 0.104776 |
|           | market_value_in_eur                      | 1.000000 |
|           | remaining_contract_days                  | 0.334212 |
|           | games_total                              | 0.470027 |
|           | goals_total                              | 0.467667 |
|           | assists_total                            | 0.486171 |
|           | minutes_played_total                     | 0.471368 |
|           | goals_for_total                          | 0.545848 |
|           | goals_against_total                      | 0.392081 |
|           | clean_sheet_total                        | 0.519063 |
|           | yellow_cards_total                       | 0.352771 |
|           | red_cards_total                          | 0.115185 |
|           | games_2019                               | 0.319025 |
|           | goals_2019                               | 0.297224 |
|           | assists_2019                             | 0.303601 |
|           | minutes_played_2019                      | 0.304588 |
|           | goals_for_2019                           | 0.365819 |
|           | goals_against_2019                       | 0.268493 |
|           | clean_sheet_2019                         | 0.335958 |
|           | yellow_cards_2019                        | 0.225521 |
|           | games_2020                               | 0.361509 |
|           | goals_2020                               | 0.349266 |
|           | assists_2020                             | 0.359543 |
|           | minutes_played_2020                      | 0.357576 |
|           | goals_for_2020                           | 0.422572 |
|           | goals_against_2020                       | 0.291681 |
|           | clean_sheet_2020                         | 0.395995 |
|           | yellow_cards_2020                        | 0.238545 |
|           | games_2021                               | 0.412321 |
|           | goals_2021                               | 0.413660 |
|           | assists_2021                             | 0.427364 |
|           | minutes_played_2021                      | 0.416458 |
|           | goals_for_2021                           | 0.495140 |
|           | goals_against_2021                       | 0.338094 |
|           | clean_sheet_2021                         | 0.447354 |
|           | yellow_cards_2021                        | 0.296076 |
|           | games_2022                               | 0.434351 |
|           | goals_2022                               | 0.457946 |
|           | assists_2022                             | 0.452212 |
|           | minutes_played_2022                      | 0.446138 |
|           | goals_for_2022                           | 0.518047 |
|           | goals_against_2022                       | 0.312198 |
|           | clean_sheet_2022                         | 0.488459 |
|           | yellow_cards_2022                        | 0.314594 |
|           | games_2023                               | 0.319920 |
|           | goals_2023                               | 0.329007 |
|           | assists_2023                             | 0.273764 |
|           | minutes_played_2023                      | 0.324534 |
|           | goals_for_2023                           | 0.406962 |
|           | goals_against_2023                       | 0.223611 |
|           | clean_sheet_2023                         | 0.338500 |
|           | yellow_cards_2023                        | 0.206615 |
|           | last_season_2022                         | 0.195083 |
|           | last_season_2023                         | 0.195083 |
|           | current_club Domestic_competition_id GB1 | 0.320959 |
|           | current_club_name_Arsenal FC             | 0.160583 |
|           | current_club_name_Bayern Munich          | 0.141541 |
|           | current_club_name_Chelsea FC             | 0.148239 |
|           | current_club_name_FC Barcelona           | 0.128979 |
|           | current_club_name_Liverpool FC           | 0.138185 |
|           | current_club_name_Manchester City        | 0.210800 |
|           | current_club_name_Manchester United      | 0.118698 |
|           |  | 0.172725 |
|           | current_club_name_Real Madrid            | 0.159904 |

```
current_club_name_Tottenham Hotspur          0.106932
Name: market_value_in_eur, dtype: float64
```

Again, most categorical variables do not seem to have a high correlation with the target variable. The current domestic competition and playing for one of the larger clubs does however influence market value.

## Feature Selection

All the above features with a correlation over 0.1 will be selected into a dataset for testing along with the full datasets.

```
In [257... df8_dummies_all_fs = df8_dummies_all[['last_season_2022', 'last_season_2023', 'curr
yellow_cards_total', 'red_cards_total', 'gan
assists_2020', 'minutes_played_2020', 'goals_
clean_sheet_2021', 'yellow_cards_2021', 'game
minutes_played_2023', 'goals_for_2023', 'goa
current_club_name_Manchester City', 'current_
```

```
In [258... df8_dummies_all_fs.info()
```

```

<class 'pandas.core.frame.DataFrame'>
Index: 10168 entries, 72 to 30129
Data columns (total 65 columns):
 #   Column           Non-Null Count Dtype
 ---  -- 
 0   last_season_2022 10168 non-null  int32
 1   last_season_2023 10168 non-null  int32
 2   current_club_id  10168 non-null  int64
 3   market_value_in_eur 10168 non-null  float64
 4   remaining_contract_days 10168 non-null  int32
 5   games_total       10168 non-null  int64
 6   goals_total       10168 non-null  int64
 7   assists_total     10168 non-null  int64
 8   minutes_played_total 10168 non-null  int64
 9   goals_for_total  10168 non-null  int64
 10  goals_against_total 10168 non-null  int64
 11  clean_sheet_total 10168 non-null  int64
 12  yellow_cards_total 10168 non-null  int64
 13  red_cards_total  10168 non-null  int64
 14  games_2019        10168 non-null  int32
 15  goals_2019        10168 non-null  int32
 16  assists_2019      10168 non-null  int32
 17  minutes_played_2019 10168 non-null  int32
 18  goals_for_2019    10168 non-null  int32
 19  goals_against_2019 10168 non-null  int32
 20  clean_sheet_2019  10168 non-null  int32
 21  yellow_cards_2019 10168 non-null  int32
 22  games_2020        10168 non-null  int32
 23  goals_2020        10168 non-null  int32
 24  assists_2020      10168 non-null  int32
 25  minutes_played_2020 10168 non-null  int32
 26  goals_for_2020    10168 non-null  int32
 27  goals_against_2020 10168 non-null  int32
 28  clean_sheet_2020  10168 non-null  int32
 29  yellow_cards_2020 10168 non-null  int32
 30  games_2021        10168 non-null  int32
 31  goals_2021        10168 non-null  int32
 32  assists_2021      10168 non-null  int32
 33  minutes_played_2021 10168 non-null  int32
 34  goals_for_2021    10168 non-null  int32
 35  goals_against_2021 10168 non-null  int32
 36  clean_sheet_2021  10168 non-null  int32
 37  yellow_cards_2021 10168 non-null  int32
 38  games_2022        10168 non-null  int32
 39  goals_2022        10168 non-null  int64
 40  assists_2022      10168 non-null  int64
 41  minutes_played_2022 10168 non-null  int32
 42  goals_for_2022    10168 non-null  int32
 43  goals_against_2022 10168 non-null  int32
 44  clean_sheet_2022  10168 non-null  int64
 45  yellow_cards_2022 10168 non-null  int64
 46  games_2023        10168 non-null  int32
 47  goals_2023        10168 non-null  int32
 48  assists_2023      10168 non-null  int32
 49  minutes_played_2023 10168 non-null  int32
 50  goals_for_2023    10168 non-null  int32
 51  goals_against_2023 10168 non-null  int32
 52  clean_sheet_2023  10168 non-null  int32
 53  yellow_cards_2023 10168 non-null  int32
 54  current_club Domestic_competition_id_GB1 10168 non-null  int32
 55  current_club_name_Arsenal FC 10168 non-null  int32
 56  current_club_name_Bayern Munich 10168 non-null  int32

```

```
59 current_club_name_Liverpool FC           10168 non-null int32
60 current_club_name_Manchester City        10168 non-null int32
61 current_club_name_Manchester United       10168 non-null int32
62 current_club_name_Paris Saint-Germain    10168 non-null int32
63 current_club_name_Real Madrid            10168 non-null int32
64 current_club_name_Tottenham Hotspur       10168 non-null int32
dtypes: float64(1), int32(50), int64(14)
memory usage: 3.2 MB
```

## Scaling the Data

Both standard and robust scalers transform inputs to comparable scales. The difference lies in how they scale raw input values. Standard scaling uses mean and standard deviation. MinMax Scaler is often used as an alternative to Standard Scaler if zero mean and unit variance want to be avoided. Robust scaling uses median and interquartile range (IQR) instead. As the data contains large numbers and outliers, Robust scaler would be the preferred choice but MinMaxScaler method will also be applied for comparative purposes as it accounts for mean and unit variance.

```
In [259...]: numerical_features_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 10168 entries, 72 to 30129
Data columns (total 71 columns):
 #   Column           Non-Null Count Dtype  
 --- 
 0   current_club_id    10168 non-null  int64  
 1   height_in_cm       10168 non-null  float64 
 2   market_value_in_eur 10168 non-null  float64 
 3   age                10168 non-null  int64  
 4   remaining_contract_days 10168 non-null  int32  
 5   games_total        10168 non-null  int64  
 6   goals_total        10168 non-null  int64  
 7   assists_total      10168 non-null  int64  
 8   minutes_played_total 10168 non-null  int64  
 9   goals_for_total    10168 non-null  int64  
 10  goals_against_total 10168 non-null  int64  
 11  clean_sheet_total  10168 non-null  int64  
 12  yellow_cards_total 10168 non-null  int64  
 13  red_cards_total    10168 non-null  int64  
 14  games_2019         10168 non-null  int32  
 15  goals_2019         10168 non-null  int32  
 16  assists_2019       10168 non-null  int32  
 17  minutes_played_2019 10168 non-null  int32  
 18  goals_for_2019     10168 non-null  int32  
 19  goals_against_2019 10168 non-null  int32  
 20  clean_sheet_2019   10168 non-null  int32  
 21  yellow_cards_2019  10168 non-null  int32  
 22  red_cards_2019     10168 non-null  int32  
 23  games_2020         10168 non-null  int32  
 24  goals_2020         10168 non-null  int32  
 25  assists_2020       10168 non-null  int32  
 26  minutes_played_2020 10168 non-null  int32  
 27  goals_for_2020     10168 non-null  int32  
 28  goals_against_2020 10168 non-null  int32  
 29  clean_sheet_2020   10168 non-null  int32  
 30  yellow_cards_2020  10168 non-null  int32  
 31  red_cards_2020     10168 non-null  int32  
 32  games_2021         10168 non-null  int32  
 33  goals_2021         10168 non-null  int32  
 34  assists_2021       10168 non-null  int32  
 35  minutes_played_2021 10168 non-null  int32  
 36  goals_for_2021     10168 non-null  int32  
 37  goals_against_2021 10168 non-null  int32  
 38  clean_sheet_2021   10168 non-null  int32  
 39  yellow_cards_2021  10168 non-null  int32  
 40  red_cards_2021     10168 non-null  int32  
 41  games_2022         10168 non-null  int32  
 42  goals_2022         10168 non-null  int64  
 43  assists_2022       10168 non-null  int64  
 44  minutes_played_2022 10168 non-null  int32  
 45  goals_for_2022     10168 non-null  int32  
 46  goals_against_2022 10168 non-null  int32  
 47  clean_sheet_2022   10168 non-null  int64  
 48  yellow_cards_2022  10168 non-null  int64  
 49  red_cards_2022     10168 non-null  int64  
 50  games_2023         10168 non-null  int32  
 51  goals_2023         10168 non-null  int32  
 52  assists_2023       10168 non-null  int32  
 53  minutes_played_2023 10168 non-null  int32  
 54  goals_for_2023     10168 non-null  int32  
 55  goals_against_2023 10168 non-null  int32  
 56  clean_sheet_2023   10168 non-null  int32 
```

```

59 mins_per_goal_2023      10168 non-null float64
60 mins_per_goal_2022      10168 non-null float64
61 mins_per_goal_2021      10168 non-null float64
62 mins_per_goal_2020      10168 non-null float64
63 mins_per_goal_2019      10168 non-null float64
64 mins_per_goal_total     10168 non-null float64
65 mins_per_assist_2023    10168 non-null float64
66 mins_per_assist_2022    10168 non-null float64
67 mins_per_assist_2021    10168 non-null float64
68 mins_per_assist_2020    10168 non-null float64
69 mins_per_assist_2019    10168 non-null float64
70 mins_per_assist_total   10168 non-null float64
dtypes: float64(14), int32(41), int64(16)
memory usage: 4.0 MB

```

In [260...]: numerical\_features\_df.describe()

|              | current_club_id | height_in_cm | market_value_in_eur | age          | remaining_contract_days |
|--------------|-----------------|--------------|---------------------|--------------|-------------------------|
| <b>count</b> | 10168.000000    | 10168.000000 | 1.016800e+04        | 10168.000000 | 10168.000000            |
| <b>mean</b>  | 4836.920437     | 182.662866   | 3.884016e+06        | 25.857887    | 609.388474              |
| <b>std</b>   | 12192.771146    | 6.803874     | 9.808620e+06        | 4.861793     | 431.124416              |
| <b>min</b>   | 3.000000        | 160.000000   | 1.000000e+04        | 16.000000    | -203.000000             |
| <b>25%</b>   | 370.000000      | 178.000000   | 3.000000e+05        | 22.000000    | 193.000000              |
| <b>50%</b>   | 985.000000      | 183.000000   | 8.000000e+05        | 25.000000    | 558.000000              |
| <b>75%</b>   | 2778.000000     | 187.966463   | 2.800000e+06        | 29.000000    | 923.000000              |
| <b>max</b>   | 83678.000000    | 206.000000   | 1.800000e+08        | 43.000000    | 3115.000000             |



In [261...]: numerical\_features\_df.head()

|            | current_club_id | height_in_cm | market_value_in_eur | age | remaining_contract_days | games_tota |
|------------|-----------------|--------------|---------------------|-----|-------------------------|------------|
| <b>72</b>  | 418             | 185.0        | 25000000.0          | 36  |                         | 923        |
| <b>88</b>  | 678             | 188.0        | 200000.0            | 39  |                         | 193        |
| <b>117</b> | 367             | 180.0        | 2500000.0           | 36  |                         | 193        |
| <b>132</b> | 1095            | 184.0        | 200000.0            | 38  |                         | 609        |
| <b>135</b> | 1519            | 185.0        | 200000.0            | 37  |                         | 609        |



## MinMaxScaler

In [264...]: # Removing feature with nuerical values that are too Large for scaling  
numerical\_features\_df\_4\_midmax = numerical\_features\_df.drop(['mins\_per\_goal\_2023',

In [265...]: numerical\_features\_df\_4\_midmax.info()

```

<class 'pandas.core.frame.DataFrame'>
Index: 10168 entries, 72 to 30129
Data columns (total 59 columns):
 #   Column           Non-Null Count Dtype  
 --- 
 0   current_club_id    10168 non-null  int64  
 1   height_in_cm       10168 non-null  float64 
 2   market_value_in_eur 10168 non-null  float64 
 3   age                10168 non-null  int64  
 4   remaining_contract_days 10168 non-null  int32  
 5   games_total        10168 non-null  int64  
 6   goals_total        10168 non-null  int64  
 7   assists_total      10168 non-null  int64  
 8   minutes_played_total 10168 non-null  int64  
 9   goals_for_total    10168 non-null  int64  
 10  goals_against_total 10168 non-null  int64  
 11  clean_sheet_total  10168 non-null  int64  
 12  yellow_cards_total 10168 non-null  int64  
 13  red_cards_total    10168 non-null  int64  
 14  games_2019         10168 non-null  int32  
 15  goals_2019         10168 non-null  int32  
 16  assists_2019       10168 non-null  int32  
 17  minutes_played_2019 10168 non-null  int32  
 18  goals_for_2019     10168 non-null  int32  
 19  goals_against_2019 10168 non-null  int32  
 20  clean_sheet_2019   10168 non-null  int32  
 21  yellow_cards_2019  10168 non-null  int32  
 22  red_cards_2019     10168 non-null  int32  
 23  games_2020         10168 non-null  int32  
 24  goals_2020         10168 non-null  int32  
 25  assists_2020       10168 non-null  int32  
 26  minutes_played_2020 10168 non-null  int32  
 27  goals_for_2020     10168 non-null  int32  
 28  goals_against_2020 10168 non-null  int32  
 29  clean_sheet_2020   10168 non-null  int32  
 30  yellow_cards_2020  10168 non-null  int32  
 31  red_cards_2020     10168 non-null  int32  
 32  games_2021         10168 non-null  int32  
 33  goals_2021         10168 non-null  int32  
 34  assists_2021       10168 non-null  int32  
 35  minutes_played_2021 10168 non-null  int32  
 36  goals_for_2021     10168 non-null  int32  
 37  goals_against_2021 10168 non-null  int32  
 38  clean_sheet_2021   10168 non-null  int32  
 39  yellow_cards_2021  10168 non-null  int32  
 40  red_cards_2021     10168 non-null  int32  
 41  games_2022         10168 non-null  int32  
 42  goals_2022         10168 non-null  int64  
 43  assists_2022       10168 non-null  int64  
 44  minutes_played_2022 10168 non-null  int32  
 45  goals_for_2022     10168 non-null  int32  
 46  goals_against_2022 10168 non-null  int32  
 47  clean_sheet_2022   10168 non-null  int64  
 48  yellow_cards_2022  10168 non-null  int64  
 49  red_cards_2022     10168 non-null  int64  
 50  games_2023         10168 non-null  int32  
 51  goals_2023         10168 non-null  int32  
 52  assists_2023       10168 non-null  int32  
 53  minutes_played_2023 10168 non-null  int32  
 54  goals_for_2023     10168 non-null  int32  
 55  goals_against_2023 10168 non-null  int32  
 56  clean_sheet_2023   10168 non-null  int32  

```

```
dtypes: float64(2), int32(41), int64(16)
memory usage: 3.1 MB
```

```
In [272... numerical_features_df_4_midmax.columns
```

```
Out[272]: Index(['current_club_id', 'height_in_cm', 'market_value_in_eur', 'age',
       'remaining_contract_days', 'games_total', 'goals_total',
       'assists_total', 'minutes_played_total', 'goals_for_total',
       'goals_against_total', 'clean_sheet_total', 'yellow_cards_total',
       'red_cards_total', 'games_2019', 'goals_2019', 'assists_2019',
       'minutes_played_2019', 'goals_for_2019', 'goals_against_2019',
       'clean_sheet_2019', 'yellow_cards_2019', 'red_cards_2019', 'games_2020',
       'goals_2020', 'assists_2020', 'minutes_played_2020', 'goals_for_2020',
       'goals_against_2020', 'clean_sheet_2020', 'yellow_cards_2020',
       'red_cards_2020', 'games_2021', 'goals_2021', 'assists_2021',
       'minutes_played_2021', 'goals_for_2021', 'goals_against_2021',
       'clean_sheet_2021', 'yellow_cards_2021', 'red_cards_2021', 'games_2022',
       'goals_2022', 'assists_2022', 'minutes_played_2022', 'goals_for_2022',
       'goals_against_2022', 'clean_sheet_2022', 'yellow_cards_2022',
       'red_cards_2022', 'games_2023', 'goals_2023', 'assists_2023',
       'minutes_played_2023', 'goals_for_2023', 'goals_against_2023',
       'clean_sheet_2023', 'yellow_cards_2023', 'red_cards_2023'],
      dtype='object')
```

```
In [262... # Creating a MinMaxScaler object
mmscaler = MinMaxScaler()
```

```
In [266... # Train the MinMaxScaler model
mmscaler.fit(numerical_features_df_4_midmax)
```

```
Out[266]: ▾ MinMaxScaler
MinMaxScaler()
```

```
In [268... # Transform the data
```

```
numerical_features_df_4_midmax_scaled = mmscaler.transform(numerical_features_df_4_
```

Concatenating dateframes following scaling for numerical categories

```
In [276... numerical_features_df_4_midmax_scaled = pd.DataFrame(numerical_features_df_4_midmax
       'remaining_contract_days', 'games_total', 'goals_total',
       'assists_total', 'minutes_played_total', 'goals_for_total',
       'goals_against_total', 'clean_sheet_total', 'yellow_cards_total',
       'red_cards_total', 'games_2019', 'goals_2019', 'assists_2019',
       'minutes_played_2019', 'goals_for_2019', 'goals_against_2019',
       'clean_sheet_2019', 'yellow_cards_2019', 'red_cards_2019', 'games_2020',
       'goals_2020', 'assists_2020', 'minutes_played_2020', 'goals_for_2020',
       'goals_against_2020', 'clean_sheet_2020', 'yellow_cards_2020',
       'red_cards_2020', 'games_2021', 'goals_2021', 'assists_2021',
       'minutes_played_2021', 'goals_for_2021', 'goals_against_2021',
       'clean_sheet_2021', 'yellow_cards_2021', 'red_cards_2021', 'games_2022',
       'goals_2022', 'assists_2022', 'minutes_played_2022', 'goals_for_2022',
       'goals_against_2022', 'clean_sheet_2022', 'yellow_cards_2022',
       'red_cards_2022', 'games_2023', 'goals_2023', 'assists_2023',
       'minutes_played_2023', 'goals_for_2023', 'goals_against_2023',
       'clean_sheet_2023', 'yellow_cards_2023', 'red_cards_2023'])
```

```
In [277... df8_dummies_essential_mm_scaled = pd.concat([numerical_features_df_4_midmax_scaled,
```

```

<class 'pandas.core.frame.DataFrame'>
Index: 17124 entries, 0 to 30129
Data columns (total 86 columns):
 #   Column           Non-Null Count Dtype
 ---  -- 
 0   current_club_id    10168 non-null float64
 1   height_in_cm       10168 non-null float64
 2   market_value_in_eur 10168 non-null float64
 3   age                10168 non-null float64
 4   remaining_contract_days 10168 non-null float64
 5   games_total        10168 non-null float64
 6   goals_total         10168 non-null float64
 7   assists_total       10168 non-null float64
 8   minutes_played_total 10168 non-null float64
 9   goals_for_total    10168 non-null float64
 10  goals_against_total 10168 non-null float64
 11  clean_sheet_total  10168 non-null float64
 12  yellow_cards_total 10168 non-null float64
 13  red_cards_total    10168 non-null float64
 14  games_2019          10168 non-null float64
 15  goals_2019          10168 non-null float64
 16  assists_2019        10168 non-null float64
 17  minutes_played_2019 10168 non-null float64
 18  goals_for_2019     10168 non-null float64
 19  goals_against_2019 10168 non-null float64
 20  clean_sheet_2019   10168 non-null float64
 21  yellow_cards_2019  10168 non-null float64
 22  red_cards_2019     10168 non-null float64
 23  games_2020          10168 non-null float64
 24  goals_2020          10168 non-null float64
 25  assists_2020        10168 non-null float64
 26  minutes_played_2020 10168 non-null float64
 27  goals_for_2020      10168 non-null float64
 28  goals_against_2020 10168 non-null float64
 29  clean_sheet_2020   10168 non-null float64
 30  yellow_cards_2020  10168 non-null float64
 31  red_cards_2020      10168 non-null float64
 32  games_2021          10168 non-null float64
 33  goals_2021          10168 non-null float64
 34  assists_2021        10168 non-null float64
 35  minutes_played_2021 10168 non-null float64
 36  goals_for_2021      10168 non-null float64
 37  goals_against_2021 10168 non-null float64
 38  clean_sheet_2021   10168 non-null float64
 39  yellow_cards_2021  10168 non-null float64
 40  red_cards_2021      10168 non-null float64
 41  games_2022          10168 non-null float64
 42  goals_2022          10168 non-null float64
 43  assists_2022        10168 non-null float64
 44  minutes_played_2022 10168 non-null float64
 45  goals_for_2022      10168 non-null float64
 46  goals_against_2022 10168 non-null float64
 47  clean_sheet_2022   10168 non-null float64
 48  yellow_cards_2022  10168 non-null float64
 49  red_cards_2022      10168 non-null float64
 50  games_2023          10168 non-null float64
 51  goals_2023          10168 non-null float64
 52  assists_2023        10168 non-null float64
 53  minutes_played_2023 10168 non-null float64
 54  goals_for_2023      10168 non-null float64
 55  goals_against_2023 10168 non-null float64
 56  clean_sheet_2023   10168 non-null float64

```

```
59 last_season_2022          10168 non-null float64
60 last_season_2023          10168 non-null float64
61 sub_position_Attacking Midfield 10168 non-null float64
62 sub_position_Central Midfield 10168 non-null float64
63 sub_position_Centre-Back    10168 non-null float64
64 sub_position_Centre-Forward 10168 non-null float64
65 sub_position_Defensive Midfield 10168 non-null float64
66 sub_position_Goalkeeper     10168 non-null float64
67 sub_position_Left Midfield 10168 non-null float64
68 sub_position_Left Winger   10168 non-null float64
69 sub_position_Left-Back     10168 non-null float64
70 sub_position_Right Midfield 10168 non-null float64
71 sub_position_Right Winger   10168 non-null float64
72 sub_position_Right-Back    10168 non-null float64
73 sub_position_Second Striker 10168 non-null float64
74 position_Attack           10168 non-null float64
75 position_Defender         10168 non-null float64
76 position_Goalkeeper        10168 non-null float64
77 position_Midfield          10168 non-null float64
78 foot_both                  10168 non-null float64
79 foot_left                  10168 non-null float64
80 foot_right                 10168 non-null float64
81 age_group_Ages 22 - 26      10168 non-null float64
82 age_group_Ages 27 - 32      10168 non-null float64
83 age_group_Ages 33 - 38      10168 non-null float64
84 age_group_Ages 39 - 44      10168 non-null float64
85 age_group_Ages Under 22    10168 non-null float64

dtypes: float64(86)
memory usage: 11.4 MB
```

```
In [279...]: df8_dummies_all_mm_scaled = pd.concat([numerical_features_df_4_midmax_scaled, all_d
```

```
In [280]: df8_dummies_all_mm_scaled.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 17124 entries, 0 to 30129
Columns: 374 entries, current_club_id to current_club_name_Ümraniyespor
dtypes: float64(374)
memory usage: 49.0 MB
```

Categorical Features :['last\_season','sub\_position', 'position', 'foot',  
'current\_club\_domestic\_competition\_id', 'current\_club\_name', 'age\_group']

```
In [281]: df8_dummies_all_fs_mm_scaled = numerical_features_df_4_midmax_scaled[['current_club',
    'yellow_cards_total', 'red_cards_total', 'goals_assists_2020', 'minutes_played_2020', 'goals_clean_sheet_2021',
    'yellow_cards_2021', 'game_minutes_played_2023', 'goals_for_2023', 'goals_against_2023']]
```

```
In [285]: feature_selected_dummies = all_dummies[['last_season_2022', 'last_season_2023', 'current_club_name_Manchester City', 'current_
```

```
In [286]: df8_dummies_all_fs_mm_scaled = pd.concat([df8_dummies_all_fs_mm_scaled, feature_sel])
```

```
In [287]: df8.dummies.all().fs_mm_scaled.info()
```

```

<class 'pandas.core.frame.DataFrame'>
Index: 17124 entries, 0 to 30129
Data columns (total 65 columns):
 #   Column           Non-Null Count Dtype
 ---  -- 
 0   current_club_id    10168 non-null  float64
 1   market_value_in_eur 10168 non-null  float64
 2   remaining_contract_days 10168 non-null  float64
 3   games_total        10168 non-null  float64
 4   goals_total         10168 non-null  float64
 5   assists_total       10168 non-null  float64
 6   minutes_played_total 10168 non-null  float64
 7   goals_for_total     10168 non-null  float64
 8   goals_against_total 10168 non-null  float64
 9   clean_sheet_total   10168 non-null  float64
 10  yellow_cards_total  10168 non-null  float64
 11  red_cards_total     10168 non-null  float64
 12  games_2019          10168 non-null  float64
 13  goals_2019          10168 non-null  float64
 14  assists_2019         10168 non-null  float64
 15  minutes_played_2019 10168 non-null  float64
 16  goals_for_2019      10168 non-null  float64
 17  goals_against_2019   10168 non-null  float64
 18  clean_sheet_2019     10168 non-null  float64
 19  yellow_cards_2019    10168 non-null  float64
 20  games_2020          10168 non-null  float64
 21  goals_2020          10168 non-null  float64
 22  assists_2020         10168 non-null  float64
 23  minutes_played_2020 10168 non-null  float64
 24  goals_for_2020       10168 non-null  float64
 25  goals_against_2020   10168 non-null  float64
 26  clean_sheet_2020     10168 non-null  float64
 27  yellow_cards_2020    10168 non-null  float64
 28  games_2021          10168 non-null  float64
 29  goals_2021          10168 non-null  float64
 30  assists_2021         10168 non-null  float64
 31  minutes_played_2021 10168 non-null  float64
 32  goals_for_2021       10168 non-null  float64
 33  goals_against_2021   10168 non-null  float64
 34  clean_sheet_2021     10168 non-null  float64
 35  yellow_cards_2021    10168 non-null  float64
 36  games_2022          10168 non-null  float64
 37  goals_2022          10168 non-null  float64
 38  assists_2022         10168 non-null  float64
 39  minutes_played_2022 10168 non-null  float64
 40  goals_for_2022       10168 non-null  float64
 41  goals_against_2022   10168 non-null  float64
 42  clean_sheet_2022     10168 non-null  float64
 43  yellow_cards_2022    10168 non-null  float64
 44  games_2023          10168 non-null  float64
 45  goals_2023          10168 non-null  float64
 46  assists_2023         10168 non-null  float64
 47  minutes_played_2023 10168 non-null  float64
 48  goals_for_2023       10168 non-null  float64
 49  goals_against_2023   10168 non-null  float64
 50  clean_sheet_2023     10168 non-null  float64
 51  yellow_cards_2023    10168 non-null  float64
 52  last_season_2022     10168 non-null  float64
 53  last_season_2023     10168 non-null  float64
 54  current_club Domestic_competition_id_GB1 10168 non-null  float64
 55  current_club_name_Arsenal FC 10168 non-null  float64
 56  current_club_name_Bayern Munich 10168 non-null  float64

```

```

59 current_club_name_Liverpool FC           10168 non-null float64
60 current_club_name_Manchester City       10168 non-null float64
61 current_club_name_Manchester United     10168 non-null float64
62 current_club_name_Paris Saint-Germain   10168 non-null float64
63 current_club_name_Real Madrid          10168 non-null float64
64 current_club_name_Tottenham Hotspur      10168 non-null float64
dtypes: float64(65)
memory usage: 8.6 MB

```

In [303]: `df8_dummies_all_fs_mm_scaled.describe()`

Out[303]:

|              | <b>current_club_id</b> | <b>market_value_in_eur</b> | <b>remaining_contract_days</b> | <b>games_total</b> | <b>goals_total</b> |
|--------------|------------------------|----------------------------|--------------------------------|--------------------|--------------------|
| <b>count</b> | 10168.000000           | 10168.000000               | 10168.000000                   | 10168.000000       | 10168.000000       |
| <b>mean</b>  | 0.057770               | 0.021524                   | 0.244843                       | 0.217469           | 0.023830           |
| <b>std</b>   | 0.145716               | 0.054495                   | 0.129935                       | 0.223635           | 0.052570           |
| <b>min</b>   | 0.000000               | 0.000000                   | 0.000000                       | 0.000000           | 0.000000           |
| <b>25%</b>   | 0.004386               | 0.001611                   | 0.119349                       | 0.023256           | 0.000000           |
| <b>50%</b>   | 0.011736               | 0.004389                   | 0.229355                       | 0.139535           | 0.005208           |
| <b>75%</b>   | 0.033164               | 0.015501                   | 0.339361                       | 0.367442           | 0.020833           |
| <b>max</b>   | 1.000000               | 1.000000                   | 1.000000                       | 1.000000           | 1.000000           |



## Robust Scaler

As with the MidMaxScaler, Robust scaler does not seem to work with the features generated with large (infinity like) values. We will scale the remaining numerical features.

In [288]: `# Creating a RobustScaler object  
rbscaler = RobustScaler()`

In [291]: `# Train the RobustScaler model  
rbscaler.fit(numerical_features_df_4_midmax)`

Out[291]:

RobustScaler()

In [292...]: `# Transform the data  
numerical_features_df_rb_scaled = rbscaler.transform(numerical_features_df_4_midma`

Concatenating dateframes following robust scaling of numerical categories

In [293...]: `numerical_features_df_rb_scaled = pd.DataFrame(numerical_features_df_rb_scaled, columns=['remaining_contract_days', 'games_total', 'goals_total', 'assists_total', 'minutes_played_total', 'goals_for_total', 'goals_against_total', 'clean_sheet_total', 'yellow_cards_total', 'red_cards_total', 'games_2019', 'goals_2019', 'assists_2019', 'minutes_played_2019', 'goals_for_2019', 'goals_against_2019', 'clean_sheet_2019', 'yellow_cards_2019', 'red_cards_2019', 'games_2020', 'goals_2020', 'assists_2020', 'minutes_played_2020', 'goals_for_2020', 'goals_2021', 'assists_2021', 'sheet_2020', 'yellow_cards_2020', 'red_cards_2020', 'games_2021', 'goals_2021', 'assists_2021'])`

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js sheet\_2020', 'yellow\_cards\_2020', 'red\_cards\_2020', 'games\_2021', 'goals\_2021', 'assists\_2021',

```
'minutes_played_2021', 'goals_for_2021', 'goals_against_2021',
'clean_sheet_2021', 'yellow_cards_2021', 'red_cards_2021', 'games_2022',
'goals_2022', 'assists_2022', 'minutes_played_2022', 'goals_for_2022',
'goals_against_2022', 'clean_sheet_2022', 'yellow_cards_2022',
'red_cards_2022', 'games_2023', 'goals_2023', 'assists_2023',
'minutes_played_2023', 'goals_for_2023', 'goals_against_2023',
'clean_sheet_2023', 'yellow_cards_2023', 'red_cards_2023'])
```

In [294]: df8\_dummies\_essential\_rb\_scaled = pd.concat([numerical\_features\_df\_rb\_scaled, dummies\_essential\_rb\_scaled], axis=1)

In [295]: df8\_dummies\_essential\_rb\_scaled.info()

```

<class 'pandas.core.frame.DataFrame'>
Index: 17124 entries, 0 to 30129
Data columns (total 86 columns):
 #   Column           Non-Null Count Dtype
 ---  -- 
 0   current_club_id    10168 non-null float64
 1   height_in_cm       10168 non-null float64
 2   market_value_in_eur 10168 non-null float64
 3   age                10168 non-null float64
 4   remaining_contract_days 10168 non-null float64
 5   games_total        10168 non-null float64
 6   goals_total        10168 non-null float64
 7   assists_total      10168 non-null float64
 8   minutes_played_total 10168 non-null float64
 9   goals_for_total    10168 non-null float64
 10  goals_against_total 10168 non-null float64
 11  clean_sheet_total  10168 non-null float64
 12  yellow_cards_total 10168 non-null float64
 13  red_cards_total    10168 non-null float64
 14  games_2019         10168 non-null float64
 15  goals_2019         10168 non-null float64
 16  assists_2019       10168 non-null float64
 17  minutes_played_2019 10168 non-null float64
 18  goals_for_2019    10168 non-null float64
 19  goals_against_2019 10168 non-null float64
 20  clean_sheet_2019   10168 non-null float64
 21  yellow_cards_2019  10168 non-null float64
 22  red_cards_2019    10168 non-null float64
 23  games_2020         10168 non-null float64
 24  goals_2020         10168 non-null float64
 25  assists_2020       10168 non-null float64
 26  minutes_played_2020 10168 non-null float64
 27  goals_for_2020    10168 non-null float64
 28  goals_against_2020 10168 non-null float64
 29  clean_sheet_2020   10168 non-null float64
 30  yellow_cards_2020  10168 non-null float64
 31  red_cards_2020    10168 non-null float64
 32  games_2021         10168 non-null float64
 33  goals_2021         10168 non-null float64
 34  assists_2021       10168 non-null float64
 35  minutes_played_2021 10168 non-null float64
 36  goals_for_2021    10168 non-null float64
 37  goals_against_2021 10168 non-null float64
 38  clean_sheet_2021   10168 non-null float64
 39  yellow_cards_2021  10168 non-null float64
 40  red_cards_2021    10168 non-null float64
 41  games_2022         10168 non-null float64
 42  goals_2022         10168 non-null float64
 43  assists_2022       10168 non-null float64
 44  minutes_played_2022 10168 non-null float64
 45  goals_for_2022    10168 non-null float64
 46  goals_against_2022 10168 non-null float64
 47  clean_sheet_2022   10168 non-null float64
 48  yellow_cards_2022  10168 non-null float64
 49  red_cards_2022    10168 non-null float64
 50  games_2023         10168 non-null float64
 51  goals_2023         10168 non-null float64
 52  assists_2023       10168 non-null float64
 53  minutes_played_2023 10168 non-null float64
 54  goals_for_2023    10168 non-null float64
 55  goals_against_2023 10168 non-null float64
 56  clean_sheet_2023   10168 non-null float64

```

```
59 last_season_2022           10168 non-null float64
60 last_season_2023           10168 non-null float64
61 sub_position_Attacking Midfield 10168 non-null float64
62 sub_position_Central Midfield 10168 non-null float64
63 sub_position_Centre-Back    10168 non-null float64
64 sub_position_Centre-Forward 10168 non-null float64
65 sub_position_Defensive Midfield 10168 non-null float64
66 sub_position_Goalkeeper    10168 non-null float64
67 sub_position_Left Midfield 10168 non-null float64
68 sub_position_Left Winger   10168 non-null float64
69 sub_position_Left-Back     10168 non-null float64
70 sub_position_Right Midfield 10168 non-null float64
71 sub_position_Right Winger   10168 non-null float64
72 sub_position_Right-Back    10168 non-null float64
73 sub_position_Second Striker 10168 non-null float64
74 position_Attack           10168 non-null float64
75 position_Defender         10168 non-null float64
76 position_Goalkeeper       10168 non-null float64
77 position_Midfield         10168 non-null float64
78 foot_both                  10168 non-null float64
79 foot_left                  10168 non-null float64
80 foot_right                 10168 non-null float64
81 age_group_Ages 22 - 26     10168 non-null float64
82 age_group_Ages 27 - 32     10168 non-null float64
83 age_group_Ages 33 - 38     10168 non-null float64
84 age_group_Ages 39 - 44     10168 non-null float64
85 age_group_Ages Under 22   10168 non-null float64
dtypes: float64(86)
memory usage: 11.4 MB
```

```
In [296...]: df8_dummies_all_rb_scaled = pd.concat([numerical_features_df_rb_scaled, all_dummies])
```

```
In [297...]: df8_dummies_all_rb_scaled.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 17124 entries, 0 to 30129
Columns: 374 entries, current_club_id to current_club_name_Ümraniyespor
dtypes: float64(374)
memory usage: 49.0 MB
```

```
In [298...]: df8_dummies_all_fs_rb_scaled = numerical_features_df_rb_scaled[['current_club_id',
                                                               'yellow_cards_total', 'red_cards_total', 'gan-
                                                               'assists_2020', 'minutes_played_2020', 'goals_
                                                               'clean_sheet_2021', 'yellow_cards_2021', 'game-
                                                               'minutes_played_2023', 'goals_for_2023', 'goa-
```

```
In [300...]: df8_dummies_all_fs_rb_scaled = pd.concat([df8_dummies_all_fs_rb_scaled, feature_se])
```

```
In [301...]: df8_dummies_all_fs_rb_scaled.info()
```

```

<class 'pandas.core.frame.DataFrame'>
Index: 17124 entries, 0 to 30129
Data columns (total 65 columns):
 #   Column           Non-Null Count Dtype
 ---  -- 
 0   current_club_id    10168 non-null  float64
 1   market_value_in_eur 10168 non-null  float64
 2   remaining_contract_days 10168 non-null  float64
 3   games_total        10168 non-null  float64
 4   goals_total         10168 non-null  float64
 5   assists_total       10168 non-null  float64
 6   minutes_played_total 10168 non-null  float64
 7   goals_for_total     10168 non-null  float64
 8   goals_against_total 10168 non-null  float64
 9   clean_sheet_total   10168 non-null  float64
 10  yellow_cards_total  10168 non-null  float64
 11  red_cards_total     10168 non-null  float64
 12  games_2019          10168 non-null  float64
 13  goals_2019          10168 non-null  float64
 14  assists_2019         10168 non-null  float64
 15  minutes_played_2019 10168 non-null  float64
 16  goals_for_2019      10168 non-null  float64
 17  goals_against_2019   10168 non-null  float64
 18  clean_sheet_2019     10168 non-null  float64
 19  yellow_cards_2019    10168 non-null  float64
 20  games_2020          10168 non-null  float64
 21  goals_2020          10168 non-null  float64
 22  assists_2020         10168 non-null  float64
 23  minutes_played_2020 10168 non-null  float64
 24  goals_for_2020       10168 non-null  float64
 25  goals_against_2020   10168 non-null  float64
 26  clean_sheet_2020     10168 non-null  float64
 27  yellow_cards_2020    10168 non-null  float64
 28  games_2021          10168 non-null  float64
 29  goals_2021          10168 non-null  float64
 30  assists_2021         10168 non-null  float64
 31  minutes_played_2021 10168 non-null  float64
 32  goals_for_2021       10168 non-null  float64
 33  goals_against_2021   10168 non-null  float64
 34  clean_sheet_2021     10168 non-null  float64
 35  yellow_cards_2021    10168 non-null  float64
 36  games_2022          10168 non-null  float64
 37  goals_2022          10168 non-null  float64
 38  assists_2022         10168 non-null  float64
 39  minutes_played_2022 10168 non-null  float64
 40  goals_for_2022       10168 non-null  float64
 41  goals_against_2022   10168 non-null  float64
 42  clean_sheet_2022     10168 non-null  float64
 43  yellow_cards_2022    10168 non-null  float64
 44  games_2023          10168 non-null  float64
 45  goals_2023          10168 non-null  float64
 46  assists_2023         10168 non-null  float64
 47  minutes_played_2023 10168 non-null  float64
 48  goals_for_2023       10168 non-null  float64
 49  goals_against_2023   10168 non-null  float64
 50  clean_sheet_2023     10168 non-null  float64
 51  yellow_cards_2023    10168 non-null  float64
 52  last_season_2022     10168 non-null  float64
 53  last_season_2023     10168 non-null  float64
 54  current_club Domestic_competition_id_GB1 10168 non-null  float64
 55  current_club_name_Arsenal FC 10168 non-null  float64
 56  current_club_name_Bayern Munich 10168 non-null  float64

```

```

59 current_club_name_Liverpool FC           10168 non-null float64
60 current_club_name_Manchester City       10168 non-null float64
61 current_club_name_Manchester United     10168 non-null float64
62 current_club_name_Paris Saint-Germain   10168 non-null float64
63 current_club_name_Real Madrid          10168 non-null float64
64 current_club_name_Tottenham Hotspur     10168 non-null float64
dtypes: float64(65)
memory usage: 8.6 MB

```

In [302]: `df8_dummies_all_fs_rb_scaled.describe()`

Out[302]:

|              | <b>current_club_id</b> | <b>market_value_in_eur</b> | <b>remaining_contract_days</b> | <b>games_total</b> | <b>goals_total</b> |
|--------------|------------------------|----------------------------|--------------------------------|--------------------|--------------------|
| <b>count</b> | 10168.000000           | 10168.000000               | 10168.000000                   | 10168.000000       | 10168.000000       |
| <b>mean</b>  | 1.599635               | 1.233607                   | 0.070395                       | 0.226430           | 0.893858           |
| <b>std</b>   | 5.063443               | 3.923448                   | 0.590581                       | 0.649751           | 2.523362           |
| <b>min</b>   | -0.407807              | -0.316000                  | -1.042466                      | -0.405405          | -0.250000          |
| <b>25%</b>   | -0.255399              | -0.200000                  | -0.500000                      | -0.337838          | -0.250000          |
| <b>50%</b>   | 0.000000               | 0.000000                   | 0.000000                       | 0.000000           | 0.000000           |
| <b>75%</b>   | 0.744601               | 0.800000                   | 0.500000                       | 0.662162           | 0.750000           |
| <b>max</b>   | 34.340947              | 71.680000                  | 3.502740                       | 2.500000           | 47.750000          |

In [304]: `df8_dummies_all.info()`

```

<class 'pandas.core.frame.DataFrame'>
Index: 10168 entries, 72 to 30129
Columns: 386 entries, current_club_id to current_club_name_Ümraniyespor
dtypes: float64(14), int32(356), int64(16)
memory usage: 16.5 MB

```

## Saving all dataframes to csv following preprocessing

In [305]:

```

# NO SCALING

# One-hot Label encoded but no scaling or feature selection
df8_dummies_all.to_csv('df8_dummies_all.csv', index=False)

# MINMAX SCALER

# Only Numerical Features and Robust Scaling
numerical_features_df_4_midmax_scaled.to_csv('numerical_features_df_4_midmax_scaled.csv', index=False)

# One-hot Label encoded on select Categorical features(last_season, foot, position,
# and MinMax Scaling NO Feature Selection
df8_dummies_essential_mm_scaled.to_csv('df8_dummies_essential_mm_scaled.csv', index=False)

# One-hot Label encoded on ALL Categorical features and MinMax Scaling, NO Feature Selection
df8_dummies_all_mm_scaled.to_csv('df8_dummies_all_mm_scaled.csv', index=False)

# One-hot Label encoded on ALL Categorical features with Feature Selection and MinMax Scaling
df8_dummies_all_fs_mm_scaled.to_csv('df8_dummies_all_fs_mm_scaled.csv', index=False)

```

```

# Only Numerical Features and Robust Scaling
numerical_features_df_rb_scaled.to_csv('numerical_features_df_rb_scaled.csv', index=False)

# One-hot Label encoded on select Categorical features(last_season, foot, position,
# and Robust Scaling NO Feature Selection
df8_dummies_essential_rb_scaled.to_csv('df8_dummies_essential_rb_scaled.csv', index=False)

# One-hot Label encoded on ALL Categorical features and Robust Scaling, NO Feature Selection
df8_dummies_all_rb_scaled.to_csv('df8_dummies_all_rb_scaled.csv', index=False)

# One-hot Label encoded on ALL Categorical features with Feature Selection and Robust Scaling
df8_dummies_all_fs_rb_scaled.to_csv('df8_dummies_all_fs_rb_scaled.csv', index=False)

```

## References

Iqbal, Dr.M. (n.d.). CCT College Dublin: Log in to the site. [online] moodle.cct.ie. Available at: <https://moodle.cct.ie/course/view.php?id=2625> [Accessed 17 Nov. 2023].

MüllerA.C. and Guido, S. (2017). Introduction to machine learning with Python : a guide for data scientists. Sebastopol, CA: O'Reilly Media, Inc.

Scikit-learn.org. (2018). Compare the effect of different scalers on data with outliers — scikit-learn 0.20.3 documentation. [online] Available at: [https://scikit-learn.org/stable/auto\\_examples/preprocessing/plot\\_all\\_scaling.html](https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html).

seaborn.pydata.org. (n.d.). User guide and tutorial — seaborn 0.13.0 documentation. [online] Available at: <https://seaborn.pydata.org/tutorial>.

Ahmed, T. (n.d.). CCT College Dublin: Log in to the site. [online] moodle.cct.ie. Available at: <https://moodle.cct.ie/course/view.php?id=2637> [Accessed 20 Dec. 2023].

Nalcin, S. (2022). StandardScaler vs. MinMaxScaler vs. RobustScaler: Which one to use for your next ML project? [online] Medium. Available at: <https://medium.com/@onersarpnalcin/standardscaler-vs-minmaxscaler-vs-robustscaler-which-one-to-use-for-your-next-ml-project-ae5b44f571b9>.