In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [3]:
```
appear = pd.read_excel("appearances.xlsx")
event =  pd.read_excel("game_events.xlsx")
lineup =  pd.read_excel("game_lineups.xlsx")
games =  pd.read_excel("games.xlsx")
players = pd.read_excel("players.xlsx")
```

In [4]:
```
appear
```

Out[4]:

	appearance_id	game_id	player_id	date	player_name	competition_id	yellow_cards
0	2224728_119169	2224728	119169	2012-07-13	Aron Johannsson	DK1	0
1	2224732_161244	2224732	161244	2012-07-14	Conor O'Brien	DK1	0
2	2224729_39467	2224729	39467	2012-07-15	Clarence Goodson	DK1	0
3	2232104_119169	2232104	119169	2012-07-19	Aron Johannsson	ELQ	0
4	2219794_39475	2219794	39475	2012-07-22	Sacha Kljestan	BESC	0
...	...	...	...	...	...	...	...
3563	3415291_537467	3415291	537467	2020-09-26	Joseph Efford	BE1	0
3564	3415296_367423	3415296	367423	2020-09-26	Chris Durkin	BE1	1
3565	3431983_478940	3431983	478940	2020-09-26	Reggie Cannon	PO1	0
3566	3450575_361104	3450575	361104	2020-09-26	Sergino Dest	NL1	0
3567	3412904_124732	3412904	124732	2020-09-27	John Anthony Brooks	L1	1

3568 rows × 11 columns

In [5]:
```
event
```

Out[5]:

	game_event_id	date	game_id	minute	type	player_id	de
0	c6a3c088ed8a38d4ce074dd73b20d3da	2012-08-19	2221641	62	Substitutions	1335	
1	02d605a5c2dc4f9a6721daa583fa5405	2012-08-26	2222536	54	Cards	1321	c
2	b56c2e2e087cddb3cfe9e3d340975df9	2012-11-18	2222707	79	Substitutions	104203	
3	4a15d1fff4f476f48bb60092c61641d5	2012-11-23	2222721	72	Substitutions	104203	
4	daa97877f7edf2fda885b411d7197921	2013-05-17	2222782	63	Goals	104203	fo th
...	...	...	...	...	...	...	
1844	4acebccbc824e45d51045d8c5e164341	2023-10-31	4194147	3	Goals	355369	, L To
1845	159ad5633cf9d7c3a97b593efb6c3269	2023-10-31	4194147	9	Goals	355369	fo To
1846	daa31f19aab26eec375884aef73c73b3	2023-10-31	4194147	65	Substitutions	355369	
1847	869a13060604e769290dafe0b1f14483	2023-11-01	4194152	70	Substitutions	504215	
1848	6b7ed06f13c1f67f0d7359c310f20ba0	2023-10-31	4194154	76	Substitutions	103064	

1849 rows × 9 columns

In [6]:
```
lineup
```

Out[6]:

	game_lineups_id	game_id	type	number	player_id	player_n
0	f2570d1504fc02f4b6c7608e8dcf89a3	4087925	substitutes	34	242284	E Hol
1	f5f0da93ea8e1d8bdd799658e7c8f7cb	4087928	starting_lineup	13	145466	Tim R
2	31a4d12ec23d604779d909d26c1b5410	4087929	substitutes	26	578539	(Rich
3	776dcbef98651450db76723cb7e3b4df	4087935	substitutes	26	578539	(Rich
4	6a35ef7495303f29e7f85dbd54547fb1	4087936	starting_lineup	13	145466	Tim R
...	...	...	...	...	...	
214	ec3d266094f99ca0a8847de827e37105	4194152	starting_lineup	7	504215	Giov R
215	667840cda9bdf3b0344b8e99b306cf38	4194152	starting_lineup	23	124732	John Ant Br
216	1c5d2f60ee777760f8a757aa10c42bb1	4194154	starting_lineup	13	103064	Terr I
217	99032084fd00ffbfee52c541a9f960ab	4204000	substitutes	14	315762	Luca .
218	5c9eaf6ebb621d43a0d6fd6a9e607ef9	4220942	substitutes	13	145466	Tim R

219 rows × 8 columns

In [7]:
```
games
```

Out[7]:

	game_id	competition_id	season	round	date	home_club_goals	away_club_goals
0	2222734	RU1	2012	19. Matchday	2012-12-07	2	3
1	2224572	DK1	2012	3. Matchday	2012-07-28	1	2
2	2224628	DK1	2012	22. Matchday	2013-03-08	2	0
3	2224655	DK1	2012	22. Matchday	2013-03-10	0	3
4	2224729	DK1	2012	1. Matchday	2012-07-15	0	1
...	...	...	...	...	...	...	...
3263	3296153	NLP	2019	Round of 16	2020-01-22	7	0
3264	2875216	DK1	2017	19. Matchday	2017-12-10	3	2
3265	3099247	BE1	2018	11. Matchday	2018-10-20	3	1
3266	2872273	GB1	2017	16. Matchday	2017-12-09	5	1
3267	2517322	FAC	2014	Third Round	2015-01-04	3	1

3268 rows × 20 columns

In [8]:
```
players
```

Out[8]:

	player_id	name	last_season	current_club_id	player_code	country_of_birth	date_of
0	124732	John Anthony Brooks	2023	533	john-anthony-brooks	Germany	1993
1	223047	Emerson Hyndman	2018	903	emerson-hyndman	United States	1996
2	307781	Lynden Gooch	2016	289	lynden-gooch	United States	1995
3	370846	Timothy Weah	2023	506	timothy-weah	United States	2000
4	484756	Djordje Mihailovic	2023	1090	djordje-mihailovic	United States	1998
...	...	...	...	...	...	...	
147	273570	Desevio Payne	2020	1283	desevio-payne	United States	1995
148	111783	Alejandro Bedoya	2015	995	alejandro-bedoya	United States	1987
149	160670	Joe Gyau	2014	16	joe-gyau	United States	1992
150	3476	Brad Friedel	2014	148	brad-friedel	United States	1971
151	504215	Giovanni Reyna	2023	16	giovanni-reyna	England	2002

152 rows × 15 columns

In [9]:
```
#merging the datasets
```

In [10]:
```
df_event =  event.merge(games, on= "game_id", how = "left")
```

In [11]:
```
df_event = df_event.merge(players, on = "player_id", how="left")
```

In [12]:
```
df_appear = appear.merge(games, on = "game_id", how="left")
```

In [13]:
```
df_appear= df_appear.merge(players, on="player_id", how="left")
```

In [14]:
```
df_lineup = lineup.merge(games, on="game_id", how="left")
```

In [15]:
```
df_lineup = df_lineup.merge(players, on= "player_id", how = "left")
```

In [16]:
```
event.shape
```

Out[16]:
```
(1849, 9)
```

In [17]:
```
games.shape
```

Out[17]: (3268, 20)

In [18]:
```
df_event
```

Out[18]:

	game_event_id	date_x	game_id	minute	type	player_id	d
0	c6a3c088ed8a38d4ce074dd73b20d3da	2012-08-19	2221641	62	Substitutions	1335	
1	02d605a5c2dc4f9a6721daa583fa5405	2012-08-26	2222536	54	Cards	1321	
2	b56c2e2e087cddb3cfe9e3d340975df9	2012-11-18	2222707	79	Substitutions	104203	
3	4a15d1fff4f476f48bb60092c61641d5	2012-11-23	2222721	72	Substitutions	104203	
4	daa97877f7edf2fda885b411d7197921	2013-05-17	2222782	63	Goals	104203	f t
...	...	...	...	...	...	...	
1844	4acebccbc824e45d51045d8c5e164341	2023-10-31	4194147	3	Goals	355369	, T
1845	159ad5633cf9d7c3a97b593efb6c3269	2023-10-31	4194147	9	Goals	355369	f T
1846	daa31f19aab26eec375884aef73c73b3	2023-10-31	4194147	65	Substitutions	355369	
1847	869a13060604e769290dafe0b1f14483	2023-11-01	4194152	70	Substitutions	504215	
1848	6b7ed06f13c1f67f0d7359c310f20ba0	2023-10-31	4194154	76	Substitutions	103064	

1849 rows × 42 columns

In [16]:
```
df_appear
```

Out[16]:

	appearance_id	game_id	player_id	date_x	player_name	competition_id_x	yellow_
0	2224728_119169	2224728	119169	2012-07-13	Aron Johannsson	DK1	
1	2224732_161244	2224732	161244	2012-07-14	Conor O'Brien	DK1	
2	2224729_39467	2224729	39467	2012-07-15	Clarence Goodson	DK1	
3	2232104_119169	2232104	119169	2012-07-19	Aron Johannsson	ELQ	
4	2219794_39475	2219794	39475	2012-07-22	Sacha Kljestan	BESC	
...	...	...	...	...	...	...	
3563	3415291_537467	3415291	537467	2020-09-26	Joseph Efford	BE1	
3564	3415296_367423	3415296	367423	2020-	Chris Durkin	BE1	

In [19]:
```
df_lineup
```

Out[19]:

	game_lineups_id	game_id	type	number	player_id	player_n
0	f2570d1504fc02f4b6c7608e8dcf89a3	4087925	substitutes	34	242284	E Hol
1	f5f0da93ea8e1d8bdd799658e7c8f7cb	4087928	starting_lineup	13	145466	Tim R
2	31a4d12ec23d604779d909d26c1b5410	4087929	substitutes	26	578539	(Rich
3	776dcbef98651450db76723cb7e3b4df	4087935	substitutes	26	578539	(Rich
4	6a35ef7495303f29e7f85dbd54547fb1	4087936	starting_lineup	13	145466	Tim R
...	...	...	...	...	...	
214	ec3d266094f99ca0a8847de827e37105	4194152	starting_lineup	7	504215	Giov R
215	667840cda9bdf3b0344b8e99b306cf38	4194152	starting_lineup	23	124732	John Ant Br
216	1c5d2f60ee777760f8a757aa10c42bb1	4194154	starting_lineup	13	103064	Terr I
217	99032084fd00ffbfee52c541a9f960ab	4204000	substitutes	14	315762	Luca .
218	5c9eaf6ebb621d43a0d6fd6a9e607ef9	4220942	substitutes	13	145466	Tim R

219 rows × 41 columns

In [20]:
```
df_lineup.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 219 entries, 0 to 218
Data columns (total 41 columns):
 #   Column                       Non-Null Count  Dtype         
---  ------                       --------------  -----         
 0   game_lineups_id              219 non-null    object        
 1   game_id                      219 non-null    int64         
 2   type                         219 non-null    object        
 3   number                       219 non-null    int64         
 4   player_id                    219 non-null    int64         
 5   player_name                  219 non-null    object        
 6   team_captain                 219 non-null    int64         
 7   position_x                   219 non-null    object        
 8   competition_id               0 non-null      object        
 9   season                       0 non-null      float64       
 10  round                        0 non-null      object        
 11  date                         0 non-null      datetime64[ns]
 12  home_club_goals              0 non-null      float64       
 13  away_club_goals              0 non-null      float64       
 14  home_club_position           0 non-null      float64       
 15  away_club_position           0 non-null      float64       
 16  home_club_manager_name       0 non-null      object        
 17  away_club_manager_name       0 non-null      object        
 18  stadium                      0 non-null      object        
 19  attendance                   0 non-null      float64       
 20  referee                      0 non-null      object        
 21  home_club_formation          0 non-null      float64       
 22  away_club_formation          0 non-null      float64       
 23  home_club_name               0 non-null      object        
 24  away_club_name               0 non-null      object        
 25  aggregate                    0 non-null      object        
 26  competition_type             0 non-null      object        
 27  name                         219 non-null    object        
 28  last_season                  219 non-null    int64         
 29  current_club_id              219 non-null    int64         
 30  player_code                  219 non-null    object        
 31  country_of_birth             219 non-null    object        
 32  date_of_birth                219 non-null    datetime64[ns]
 33  sub_position                 219 non-null    object        
 34  position_y                   219 non-null    object        
 35  foot                         219 non-null    object        
 36  height_in_cm                 219 non-null    float64       
 37  market_value_in_eur          218 non-null    float64       
 38  highest_market_value_in_eur  219 non-null    float64       
 39  contract_expiration_date     210 non-null    datetime64[ns]
 40  agent_name                   152 non-null    object        
dtypes: datetime64[ns](3), float64(11), int64(6), object(21)
memory usage: 70.3+ KB
```

In [21]:
```
df_event.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1849 entries, 0 to 1848
Data columns (total 42 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   game_event_id                 1849 non-null   object
 1   date_x                        1849 non-null   datetime64[ns]
 2   game_id                       1849 non-null   int64
 3   minute                        1849 non-null   int64
 4   type                          1849 non-null   object
 5   player_id                     1849 non-null   int64
 6   description                   916 non-null    object
 7   player_in_id                  1155 non-null   float64
 8   player_assist_id              140 non-null    float64
 9   competition_id                1126 non-null   object
 10  season                        1126 non-null   float64
 11  round                         1126 non-null   object
 12  date_y                        1126 non-null   datetime64[ns]
 13  home_club_goals               1126 non-null   float64
 14  away_club_goals               1126 non-null   float64
 15  home_club_position            927 non-null    float64
 16  away_club_position            927 non-null    float64
 17  home_club_manager_name        1122 non-null   object
 18  away_club_manager_name        1122 non-null   object
 19  stadium                       1126 non-null   object
 20  attendance                    1080 non-null   float64
 21  referee                       1124 non-null   object
 22  home_club_formation           0 non-null      float64
 23  away_club_formation           0 non-null      float64
 24  home_club_name                1072 non-null   object
 25  away_club_name                1109 non-null   object
 26  aggregate                     1126 non-null   object
 27  competition_type              1126 non-null   object
 28  name                          1849 non-null   object
 29  last_season                   1849 non-null   int64
 30  current_club_id               1849 non-null   int64
 31  player_code                   1849 non-null   object
 32  country_of_birth              1849 non-null   object
 33  date_of_birth                 1849 non-null   datetime64[ns]
 34  sub_position                  1849 non-null   object
 35  position                      1849 non-null   object
 36  foot                          1824 non-null   object
 37  height_in_cm                  1833 non-null   float64
 38  market_value_in_eur           1471 non-null   float64
 39  highest_market_value_in_eur   1849 non-null   float64
 40  contract_expiration_date      1412 non-null   datetime64[ns]
 41  agent_name                    1350 non-null   object
dtypes: datetime64[ns](4), float64(13), int64(5), object(20)
memory usage: 606.8+ KB
```

In [22]:
```
df_appear.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3568 entries, 0 to 3567
Data columns (total 44 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   appearance_id               3568 non-null   object
 1   game_id                     3568 non-null   int64
 2   player_id                   3568 non-null   int64
 3   date_x                      3568 non-null   datetime64[ns]
 4   player_name                 3568 non-null   object
 5   competition_id_x            3568 non-null   object
 6   yellow_cards                3568 non-null   int64
 7   red_cards                   3568 non-null   int64
 8   goals                       3568 non-null   int64
 9   assists                     3568 non-null   int64
 10  minutes_played              3568 non-null   int64
 11  competition_id_y            3568 non-null   object
 12  season                      3568 non-null   int64
 13  round                       3568 non-null   object
 14  date_y                      3568 non-null   datetime64[ns]
 15  home_club_goals             3568 non-null   int64
 16  away_club_goals             3568 non-null   int64
 17  home_club_position          3065 non-null   float64
 18  away_club_position          3065 non-null   float64
 19  home_club_manager_name      3563 non-null   object
 20  away_club_manager_name      3563 non-null   object
 21  stadium                     3568 non-null   object
 22  attendance                  3435 non-null   float64
 23  referee                     3565 non-null   object
 24  home_club_formation         0 non-null      float64
 25  away_club_formation         0 non-null      float64
 26  home_club_name              3455 non-null   object
 27  away_club_name              3503 non-null   object
 28  aggregate                   3568 non-null   object
 29  competition_type            3568 non-null   object
 30  name                        3568 non-null   object
 31  last_season                 3568 non-null   int64
 32  current_club_id             3568 non-null   int64
 33  player_code                 3568 non-null   object
 34  country_of_birth            3568 non-null   object
 35  date_of_birth               3568 non-null   datetime64[ns]
 36  sub_position                3568 non-null   object
 37  position                    3568 non-null   object
 38  foot                        3478 non-null   object
 39  height_in_cm                3517 non-null   float64
 40  market_value_in_eur         2348 non-null   float64
 41  highest_market_value_in_eur  3568 non-null   float64
 42  contract_expiration_date    2220 non-null   datetime64[ns]
 43  agent_name                  2547 non-null   object
dtypes: datetime64[ns](4), float64(8), int64(12), object(20)
memory usage: 1.2+ MB
```

In [23]:
```
df_appear.isnull().sum()
```

Out[23]:
```
appearance_id                  0
game_id                        0
player_id                      0
date_x                         0
player_name                    0
competition_id_x               0
yellow_cards                   0
red_cards                      0
goals                          0
assists                        0
minutes_played                 0
competition_id_y               0
season                         0
round                          0
date_y                         0
home_club_goals                0
away_club_goals                0
home_club_position           503
away_club_position           503
home_club_manager_name         5
away_club_manager_name         5
stadium                        0
attendance                   133
referee                        3
home_club_formation         3568
away_club_formation         3568
home_club_name               113
away_club_name                65
aggregate                      0
competition_type               0
name                           0
last_season                    0
current_club_id                0
player_code                    0
country_of_birth               0
date_of_birth                  0
sub_position                   0
position                       0
foot                          90
height_in_cm                  51
market_value_in_eur         1220
highest_market_value_in_eur    0
contract_expiration_date    1348
agent_name                  1021
dtype: int64
```

In [24]:
```
df_event.isnull().sum()
```

Out[24]:
```
game_event_id                  0
date_x                         0
game_id                        0
minute                         0
type                           0
player_id                      0
description                  933
player_in_id                 694
player_assist_id            1709
competition_id               723
season                       723
round                        723
date_y                       723
home_club_goals              723
away_club_goals              723
home_club_position           922
away_club_position           922
home_club_manager_name       727
away_club_manager_name       727
stadium                      723
attendance                   769
referee                      725
home_club_formation         1849
away_club_formation         1849
home_club_name               777
away_club_name               740
aggregate                    723
competition_type             723
name                           0
last_season                    0
current_club_id                0
player_code                    0
country_of_birth               0
date_of_birth                  0
sub_position                   0
position                       0
foot                          25
height_in_cm                  16
market_value_in_eur          378
highest_market_value_in_eur    0
contract_expiration_date     437
agent_name                   499
dtype: int64
```

In [25]:
```
df_lineup.isnull().sum()
```

Out[25]:
```
game_lineups_id                0
game_id                        0
type                           0
number                         0
player_id                      0
player_name                    0
team_captain                   0
position_x                     0
competition_id               219
season                       219
round                        219
date                         219
home_club_goals              219
away_club_goals              219
home_club_position           219
away_club_position           219
home_club_manager_name       219
away_club_manager_name       219
stadium                      219
attendance                   219
referee                      219
home_club_formation          219
away_club_formation          219
home_club_name               219
away_club_name               219
aggregate                    219
competition_type             219
name                           0
last_season                    0
current_club_id                0
player_code                    0
country_of_birth               0
date_of_birth                  0
sub_position                   0
position_y                     0
foot                           0
height_in_cm                   0
market_value_in_eur            1
highest_market_value_in_eur    0
contract_expiration_date       9
agent_name                    67
dtype: int64
```

level 0 analysis on df_event

In [23]:
```
df_event
```

...

In [ ]:
```

```

In [26]:
```
df_event.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1849 entries, 0 to 1848
Data columns (total 42 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   game_event_id               1849 non-null   object
 1   date_x                      1849 non-null   datetime64[ns]
 2   game_id                     1849 non-null   int64
 3   minute                      1849 non-null   int64
 4   type                        1849 non-null   object
 5   player_id                   1849 non-null   int64
 6   description                 916 non-null    object
 7   player_in_id                1155 non-null   float64
 8   player_assist_id            140 non-null    float64
 9   competition_id              1126 non-null   object
 10  season                      1126 non-null   float64
 11  round                       1126 non-null   object
 12  date_y                      1126 non-null   datetime64[ns]
 13  home_club_goals             1126 non-null   float64
```

In [ ]:
```
```

In [28]:
```
#Separating the data into categorical and continuous
```

In [29]:
```
def seperate_data_types(df_event):
    categorical = []
    continuous = []
    for column in df_event.columns:
        if df_event[column].nunique() < 500:
            categorical.append(column)
        else:
            continuous.append(column)

    return categorical, continuous

categorical, continuous = seperate_data_types(df_event)

#Tabulate is a package used to print the list, dict or any data sets in a pr

from tabulate import tabulate
table = [categorical, continuous]
print(tabulate({"Categorical":categorical, "Continuous":continuous}, headers
```

```
categorical                 continuous
--------------------------  -------------
minute                      game_event_id
type                        date_x
player_id                   game_id
description                 player_in_id
player_assist_id            date_y
competition_id              attendance
season
round
home_club_goals
away_club_goals
home_club_position
away_club_position
home_club_manager_name
away_club_manager_name
stadium
referee
home_club_formation
away_club_formation
home_club_name
away_club_name
aggregate
competition_type
name
last_season
current_club_id
player_code
country_of_birth
date_of_birth
sub_position
position
foot
height_in_cm
market_value_in_eur
highest_market_value_in_eur
contract_expiration_date
agent_name
```

In [30]:
```
def info_cat(col):           #categorical
    print(col)
    print("Unique values:", df_event[col].unique())
    print("Mode(most repeated values):", df_event[col].mode()[0])
    print("Missing values:",df_event[col].isnull().sum())
```

In [31]:
```
info_cat("type")
```

```
type
Unique values: ['Substitutions' 'Cards' 'Goals']
Mode(most repeated values): Substitutions
Missing values: 0
```

In [32]:
```
info_cat("minute")
```

```
minute
Unique values: [ 62  54  79  72  63  90  65  86  56  32  33  36  48  68  4
5  12  52  11
  71  82  60  73  38  75  78  85  25  24  74  66   6  84  27  20  81  35
  44  28  37  67  46  50  69   9  80  53  88  77  39  57  58  55  42  29
  70  59  83  87  41  76  43  51  16  13  40  64   2  34   4  22  89  14
  23  49  21  31  17  61  10   5  91 105  18   1  30  26   7  19  -1   8
 108 101  47  97  15   3  98 106 102 110 103 100]
Mode(most repeated values): 90
Missing values: 0
```

In [36]:
```
info_cat("stadium")
```

```
stadium
Unique values: ['Stadion an der Kreuzeiche' 'Heinz-von-Heiden-Arena' 'Di
namo' 'Olimp-2'
 'Zvezda' 'Central Stadium  Kazan' 'Brondby Stadium' 'Sydbank Park'
 'Nordstern Arena Horsens' 'Parken' 'Ceres Park' 'MASCOT PARK'
 'Right to Dream Park' 'Aalborg Portland Park' 'MCH Arena'
 'Cepheus Park Randers' 'Nature Energy Park' 'Olympisch Stadion'
 'Constant Vanden Stock Stadion' 'Jan-Breydel-Stadion'
 'Jules Ottenstadion' "St James' Park" 'Villa Park' "St Mary's Stadium"
 'White Hart Lane' 'Old Trafford' 'Craven Cottage' 'Stadium of Light'
 'bet365 Stadium' 'Goodison Park' 'Veltins-Arena' 'PreZero Arena'
 'MERKUR SPIEL-ARENA' 'Allianz Arena' 'MHPArena Stuttgart'
 'Dreisamstadion' 'Max-Morlock-Stadion' 'BayArena' 'Mewa Arena'
 'Volksparkstadion' 'Wohninvest-Weserstadion' 'WWK ARENA'
 'Volkswagen Arena' 'Stadion im Borussia-Park' 'AFAS Stadion' 'Kamil Oca
k'
 'Bursa ataturk Stadyumu' 'Tengiz Burjanadze Stadium' nan 'Aukstaitija'
 'Emirates Stadium' 'Johan Cruijff ArenA' 'De Grolsch Veste'
 'Bingoal Stadion' 'MAC PARK stadium' 'Covebo Stadion - De Koel'
```

In [ ]:
```
```

In [35]:
```
info_cat("player_id")
```

```
player_id
Unique values: [  1335   1321 104203  39467 161244 119169  39897  46472  3
9475  39471
  61575  27577  31642  38383  49723  31041  39378  50987 103064  30593
 111783  12517 103952  72519   3476 124732 140069   4267  73427 131171
 175996 161204 209019 255916 245893 265671  58500 315779 355369 273570
 228370 176850 307781  94167 131183 223047 451860  24637 332697 190269
 381187 542776 336168 336160 537467 370846 145466 393325 332705 411302
 483047 544931 482493 282199 354613 504215 361104 465104 367423 125698
 578539 511815 478940 315762 242284 341049 605498]
Mode(most repeated values): 315779
Missing values: 0
```

In [34]:
```
info_cat("description")
```

```
description
Unique values: [', Not reported' '1. Yellow card  , Foul' ', Tactical'
 ', Right-footed shot, 1. Goal of the Season Assist: , Pass, 2. Assist o
f the Season'
 ', Right-footed shot, 2. Goal of the Season Assist: , Pass, 1. Assist o
f the Season'
 '4. Yellow card  , Professional foul'
 ', Long distance kick, 2. Goal of the Season Assist:'
 ', Left-footed shot, 2. Goal of the Season Assist: , Cross, 1. Assist o
f the Season'
 ', Tap-in, 3. Goal of the Season Assist: , Pass, 2. Assist of the Seaso
n'
 ', Tap-in, 4. Goal of the Season Assist: , Pass, 3. Assist of the Seaso
n'
 ', Penalty, 5. Goal of the Season Assist: , Fouled player'
 ', Left-footed shot, 3. Goal of the Season Assist: , Pass, 2. Assist of
the Season'
 ', Right-footed shot, 10. Goal of the Season Assist: , Pass, 3. Assist
of the Season'
```

In [37]:
```
info_cat("referee")
```

```
referee
Unique values: ['Benjamin Cortus' 'Dr. Felix Brych' 'Timur Arslanbekov'
'Sergey Karasev'
 'Aleksey Nikolaev' 'Aleksey Matyunin' 'Peter Rasmussen' 'Henning Jense
n'
 'Mads-Kristoffer Kristoffersen' 'Kenn Hansen' 'Michael Svendsen'
 'Anders Poulsen' 'Claus Bo Larsen' 'Jakob Kehlet' 'Michael Tykgaard'
 'Lars Christoffersen' 'Jens Maae' 'Michael Johansen' 'Henrik Kragh'
 'Anders Johansen' 'Joeri van de Velde' 'Luc Wouters' 'Claude Bourdouxh
e'
 'Serge Gumienny' 'Jerome Efong Nzolo' 'Jonathan Lardot' 'Lee Probert'
 'Lee Mason' 'Jonathan Moss' 'Phil Dowd' 'Chris Foy' 'Neil Swarbrick'
 'Mike Dean' 'Mike Jones' 'Martin Atkinson' 'Michael Oliver' 'Howard Web
b'
 'Andre Marriner' 'Kevin Friend' 'Mark Clattenburg' 'Daniel Siebert'
 'Markus Schmidt' 'Peter Gagelmann' 'Tobias Stieler' 'Peter Sippel'
 'Felix Zwayer' 'Marco Fritz' 'Knut Kircher' 'Michael Weiner'
 'Dr. Jochen Drees' 'Florian Meyer' 'gunter perl' 'Christian Dingert'
 'Wolfgang Stark' 'Manuel Graofe' 'Guido Winkmann' 'Deniz Aytekin'
```

In [38]:
```
info_cat("player_assist_id")
```

```
player_assist_id
Unique values: [    nan  15077.  98539.  37138.  23160.  89208.  60312.  6
8948.  15243.
   5892.  69646.  23538.  14933.  39381.   3875.   3785.  16649.   1986.
 111881.   5636.  82009. 122421.   2925.  31952.  93081.   9602.  25427.
   5017.  36391.  56534.  89231. 122011.  80709.  79121.  42004.  13465.
  74590.  14221.  37592.  21567.   7782.  45646. 132014.  43250.  37013.
  40478.  16474.  74294.  22467. 129588.  74300.  66660.   9593.  51152.
 111376.  47065. 164771.   5544.  41384.  61644.  26399. 125250. 134999.
 129554.  25118.  77001. 103064.  68864.  30700.  18297.  75921. 106987.
  49239.  53436.  49589. 361575. 416809. 116088. 467434. 524285.  82442.
 202886. 357164. 503991. 542099. 284010. 629588.]
Mode(most repeated values): 36391.0
Missing values: 1709
```

In [39]:
```
info_cat("competition_id")
```

```
competition_id
Unique values: ['DFB' 'L1' 'RU1' 'DK1' 'BE1' 'GB1' 'ELQ' 'NLP' 'TR1' nan
'CLQ' 'NL1'
 'IT1' 'CL' 'EL' 'DKP' 'CDR' 'CIT' 'FAC' 'NLSC' 'FR1' 'UKR1' 'UKRP' 'SC1'
 'SFA' 'DFL' 'ES1' 'FRCH' 'USC']
Mode(most repeated values): L1
Missing values: 723
```

In [40]:
```
info_cat("season")
```

```
season
Unique values: [2012.   nan 2013. 2014. 2015. 2016. 2017. 2018. 2019. 202
0.]
Mode(most repeated values): 2012.0
Missing values: 723
```

In [41]:
```
info_cat("round")
```

```
round
Unique values: ['First Round' '1. Matchday' '16. Matchday' '17. Matchday'
'29. Matchday'
 '24. Matchday' '7. Matchday' '6. Matchday' '9. Matchday' '10. Matchday'
 '8. Matchday' '2. Matchday' '3. Matchday' '4. Matchday' '5. Matchday'
 '13. Matchday' '14. Matchday' '11. Matchday' '12. Matchday'
 '23. Matchday' '19. Matchday' '18. Matchday' '20. Matchday'
 '21. Matchday' '31. Matchday' '33. Matchday' '28. Matchday'
 '25. Matchday' '15. Matchday' '27. Matchday' '30. Matchday'
 '38. Matchday' '32. Matchday' '37. Matchday' '34. Matchday'
 '36. Matchday' '22. Matchday' '26. Matchday' 'Second Round 1st leg'
 'Second Round' 'Second Round 2nd leg' nan '3rd round 2nd leg'
 'Qualifying Round 2nd leg' 'Group B' 'Group C' 'Group E' 'group J'
 'group L' 'Third Round' 'Round of 16' '4th round 1st leg'
 'last 16 1st leg' 'intermediate stage 1st leg' 'Quarter-Finals'
 'Third Round Replay' 'Fourth Round' 'Semi-Finals' 'last 16 2nd leg'
 'Final' 'Quarter-Finals 2nd leg' 'Qualifying Round 1st leg' 'Group A'
 'Group D' 'Group H' 'Group F' 'Round of 16 Replay'
 'intermediate stage 2nd leg' 'group K' '35. Matchday']
Mode(most repeated values): 7. Matchday
Missing values: 723
```

In [42]:
```
info_cat("home_club_goals")
```

```
home_club_goals
Unique values: [ 1.  2.  0.  3.  6.  4.  5. nan  8.  9.  7.]
Mode(most repeated values): 1.0
Missing values: 723
```

In [43]:
```
info_cat("away_club_goals")
```

```
away_club_goals
Unique values: [ 6.  2.  0.  1.  4.  3. nan  5.  9.  8.  7.]
Mode(most repeated values): 1.0
Missing values: 723
```

In [44]:
```
info_cat("home_club_position")
```

```
home_club_position
Unique values: [nan  9.  2. 12. 11.  4. 10.  6.  1.  3.  8.  5. 14. 16. 1
9. 17. 13. 15.
  7. 18. 20.]
Mode(most repeated values): 3.0
Missing values: 922
```

In [45]:
```
info_cat("away_club_position")
```

```
away_club_position
Unique values: [nan  8. 12. 16. 14. 11.  2.  7.  3.  9.  6. 10.  5.  4.
1. 19. 17. 15.
 13. 18. 20.]
Mode(most repeated values): 12.0
Missing values: 922
```

In [46]:
```
info_cat("home_club_formation")            #empty
```

...

In [ ]:
```
info_cat("away_club_formation")        #empty
```

In [47]:
```
info_cat("home_club_name")
```

```
home_club_name
Unique values: [nan 'Hannover 96' 'Anzhi Makhachkala ( -2022)' 'FK Rosto
v' 'Amkar Perm'
 'FC Rubin Kazan' 'Brondby IF' 'Sonderjyske Fc' 'AC Horsens'
 'F.C. Copenhagen' 'Aarhus Gymnastik Forening' 'Silkeborg Idrra Sforenin
g'
 'Football Club Nords' 'Aalborg BK' 'Fodbold Club Midtjylland'
 'Randers Fodbold Club' 'Odense Boldklub' 'Beerschot AC'
 'Royal Sporting Club Anderlecht'
 'Club Brugge Koninklijke Voetbalvereniging'
 'Koninklijke Atletiek Associatie Gent' 'Newcastle United Football Club'
 'Aston Villa Football Club' 'Southampton FC'
 'Tottenham Hotspur Football Club' 'Manchester United Football Club'
 'Fulham Football Club' 'Sunderland AFC' 'Stoke City'
 'Everton Football Club' 'FC Schalke 04'
 'TSG 1899 Hoffenheim Football-Spielbetriebs GmbH' 'Fortuna Dusseldorf'
 'FC Bayern Munich' 'Verein  Bewegungsspiele Stuttgart 1893'
 'Sport-Club Freiburg' '1.FC Nuremberg' 'Bayer 04 Leverkusen'
 '1. FSV Mainz 05' 'Hamburger SV' 'Sportverein Werder Bremen von 1899'
```

In [48]:
```
info_cat("away_club_name")
```

```
away_club_name
Unique values: ['Hannover 96' 'FC Schalke 04' 'FK Rostov' 'Mordovia Sarans
k (-2020)'
 'Esbjerg fB' 'Odense Boldklub' 'Aarhus Gymnastik Forening'
 'Fodbold Club Nords  Aalnd' 'Sonderjyske Fc' 'Fodbold Club Midtjylland'
 'Randers Fodbold Club' 'BRA NDBYERNES' 'Football Club Ka Benhavn'
 'Silkeborg IF' 'AC Horsens' 'Aalborg BK' 'Royal Sporting Club Anderlecht'
 'Club Brugge Koninklijke Voetbalvereniging'
 'Royal Charleroi Sporting Club' 'Royal Standard de Liege'
 'Aston Villa Football Club' 'Swansea City' 'Queens Park Rangers'
 'Tottenham Hotspur Football Club' 'Chelsea Football Club'
 'Wigan Athletic' 'Fulham Football Club' 'Newcastle United Football Club'
 'Liverpool Football Club' 'Stoke City' 'Sunderland AFC'
 'Everton Football Club' 'Manchester United Football Club'
 'Manchester City Football Club' 'Southampton FC' 'FC Augsburg 1907'
 '1.FC Nuremberg' 'tsg 1899 hoffenheim football spielbetriebs Gmbh'
 'FC Bayern Munich' 'eintracht frankfurt Football ag' 'Borussia Dortmund'
 'Sportverein Werder Bremen von 1899' 'bayer 04 leverkusen Football'
 'Verein Fur leibesubungen Wolfsburg'
 'Verein Fur Bewegungsspiele Stuttgart 1893 ' '1. FSV Mainz 05'
 'Hamburger SV' 'Spvgg Greuther Furth' 'Fortuna Dusseldorf' nan
 'Bursaspor' 'Elazigspor' 'Akhisarspor' 'Reading FC' 'Alkmaar Zaanstreek'
 'Heracles Almelo' 'Sportclub Heerenveen' 'Roda JC Kerkrade'
 'Rooms Katholieke Combinatie Waalwijk'
 'Stichting Betaald Voetbal Vitesse Arnhem' 'FC Groningen'
 'AFC Ajax Amsterdam' 'Football Club Utrecht'
 'Eindhovense Voetbalvereniging Philips Sport Vereniging'
 'Prins Hendrik Ende Desespereert Nimmer Combinatie Zwolle' 'Catania FC'
 'Atalanta Bergamasca Calcio S.p.a.' 'Associazione Sportiva Roma'
 'Genoa Cricket and Football Club' 'Anzhi Makhachkala ( -2022)'
 'Arsenal Football Club' 'FC Shakhtar Donetsk'
 'Societa Sportiva lazio S.P.A' 'Panathinaikos Athlitikos Omilos'
 'ma laga cf ' 'Olympique Lyonnais' 'Crystal Palace Football Club'
 'Hertha BSC' 'Eintracht Braunschweig' 'Sparta Rotterdam'
 'Viborg Fodsports Forening' 'Koninklijke Atletiek Associatie Gent'
 'Koninklijke Voetbalclub Kortrijk'
 'Yellow-Red Koninklijke Voetbalclub Mechelen' 'SK Beveren'
 'Oud-Heverlee Leuven' 'RAEC Mons (- 2015)'
 'Paris Saint-Germain Football Club' 'FC Sochaux Montb Liard'
 'Football Club de Nantes' 'Toulouse Football Club' 'Valenciennes FC'
 'Football Club Lorient-Bretagne Sud' 'Stade de Reims'
 'FC Girondins Bordeaux' 'EA Guingamp' 'AS Saint-Etienne' 'Norwich City'
 'PFK CSKA Moskva' 'Go Ahead Eagles' 'SC Cambuur Leeuwarden'
 'Football Club Twente' 'NAC Breda' 'Nijmegen Eendracht Combinatie'
 'Antalyaspor' 'Eskisehirspor' 'besiktas jimnastik kulubu' 'Kayserispor'
 'fenerbahce football club' 'Genclerbirligi Ankara' 'KSC Lokeren (- 2020)'
 'APS Atromitos Athinon'
 'Panthessalonikios Athlitikos Omilos Konstantinoupoliton'
 'West Ham United Football Club' 'SC Paderborn 07'
 'Borussia Verein fur Leibesubung 1900 e.V.' 'Racing Club de Lens'
 'Montpellier Herault Sport Club' 'SM Caen'
 'Villarreal Club de Football S.A.D' 'Royal Excel Mouscron (-2022)'
 'KV Oostende' 'Lierse SK (- 2018)' 'Watford FC' 'FC Ingolstadt 04'
 'ZAO FK Chornomorets Odessa' 'Vorskla Poltava'
 'Metalurg Zaporizhya (-2016)' 'PFK Stal Kamyanske (-2018)'
 'Olimpik Donetsk' 'GFC Ajaccio' 'Juventus Football Club'
 'Valencia Club de Football S. A. D' 'Sevilla Foottbol Club S.A.D'
 'FK Oleksandriya' 'Koninklijke Voetbal Club Westerlo' 'SV Zulte Waregem'
 'Lyngby Boldklubben af 1921' 'Sportverein Darmstadt 1898 e. V.'
 'RasenBallsport Leipzig' 'Sport-Club Freiburg' '1. FC Koin'
 'Inverness Caledonian Thistle FC' 'Heart of Midlothian Football Club'
```

```
 'Saint Johnstone Football Club' 'Rangers Football Club'
 'Hamilton Academical FC' 'Konyaspor' 'Sporting Clube de Braga'
 'Sport Lisboa e Benfica' 'Associazione Calcio Fiorentina'
 'Association Football Club Bournemouth' 'Koninklijke Racing Club Genk'
 'The Celtic Football Club' 'ADO Den Haag' 'Huddersfield Town' 'Hobro IK'
 'Athletic Club Bilbao' 'Stade Rennais Football Club'
 'Association sportive de Monaco Football Club' 'Naomes Olympique'
 'Leicester City' 'Brighton and Hove Albion Football Club'
 'Willem II Tilburg' 'Vendsyssel FF' 'Vejle Boldklub'
 'Hibernian Football Club' 'Kilmarnock Football Club'
 'Livingston Football Club' 'Saint Mirren Football Club'
 '1. FC Union Berlin' 'Club Atleiico Madrid S.A.D'
 'Koninklijke Sint-Truidense Voetbalvereniging'
 'Fortuna Sittardia Combinatie' 'Excelsior Rotterdam' 'FC Emmen'
 'VVV-Venlo' 'Feyenoord Rotterdam'
 "Olympique Gymnaste club Nice Cote d'azur" 'Aberdeen Football Club'
 'Wolverhampton Wanderers Football Club' 'Lille Olympique Sporting Club'
 'Allgemeine Sportvereinigung Eupen'
 'Cercle Brugge Koninklijke Sportvereniging' 'Dundee United FC']
Mode(most repeated values): Alkmaar Zaanstreek
Missing values: 740
```

In [49]:
```
info_cat("aggregate")
```

```
aggregate
Unique values: [datetime.time(1, 6) datetime.time(2, 2) datetime.time(0,
0)
 datetime.time(2, 0) datetime.time(3, 2) datetime.time(1, 1)
 datetime.time(1, 2) datetime.time(1, 4) datetime.time(2, 1)
 datetime.time(0, 4) datetime.time(0, 2) datetime.time(0, 1)
 datetime.time(3, 1) datetime.time(3, 0) datetime.time(6, 1)
 datetime.time(0, 3) datetime.time(4, 1) datetime.time(3, 3)
 datetime.time(1, 0) datetime.time(2, 3) datetime.time(2, 4)
 datetime.time(5, 3) datetime.time(4, 2) datetime.time(1, 3) nan
 datetime.time(0, 6) datetime.time(4, 0) datetime.time(6, 0)
 datetime.time(0, 5) datetime.time(5, 1) datetime.time(8, 0)
 datetime.time(2, 5) datetime.time(3, 4) datetime.time(6, 3)
 datetime.time(5, 4) datetime.time(6, 2) datetime.time(4, 3)
 datetime.time(5, 0) datetime.time(2, 6) datetime.time(1, 5)
 datetime.time(1, 9) datetime.time(9, 8) datetime.time(5, 2)
 datetime.time(4, 4) datetime.time(7, 8) datetime.time(4, 6)
 datetime.time(6, 7) datetime.time(7, 0) datetime.time(5, 7)
 datetime.time(7, 6)]
Mode(most repeated values): 01:01:00
Missing values: 723
```

In [50]:
```
info_cat("competition_type")
```

```
competition_type
Unique values: ['domestic_cup' 'domestic_league' 'international_cup' nan
'other']
Mode(most repeated values): domestic_league
Missing values: 723
```

In [51]:
```
info_cat("name")
```

```
name
Unique values: ['Steven Cherundolo' 'Jermaine Jones' 'Eugene Starikov' 'Cl
arence Goodson'
 "Conor O'Brien" 'Aron Johannsson' 'Michael Parkhurst' 'Charlie Davies'
 'Sacha Kljestan' 'Brad Guzan' 'Eric Lichaj' 'Clint Dempsey'
 'Geoff Cameron' 'Danny Williams' 'Timothy Chandler' 'Fabian Johnson'
 'Jozy Altidore' 'Maurice Edu' 'Terrence Boyd' 'Michael Bradley'
 'Alejandro Bedoya' 'Oguchi Onyewu' 'Mike Grella' 'Bobby Wood'
 'Brad Friedel' 'John Anthony Brooks' 'Babajide Ogunbiyi' 'Tim Howard'
 'Caleb Patterson-Sewell' 'Juan Agudelo' 'Sebastian Lletget'
 'Julian Green' 'Kenny Saief' 'DeAndre Yedlin' 'Matt Miazga'
 'George Fochive' 'Alfredo Morales' 'Christian Pulisic' 'Jordan'
 'Desevio Payne' 'Gedion Zelalem' 'A.J. Soares' 'Lynden Gooch'
 'Jerome Kiesewetter' 'Perry Kitchen' 'Emerson Hyndman' 'Keaton Parks'
 'Jonathan Spector' 'Weston McKennie' 'Caleb Stanko' 'Emmanuel Sabbi'
 'Jonathan Amon' 'Shaq Moore' 'Andrija Novakovich' 'Joseph Efford'
 'Timothy Weah' 'Tim Ream' 'Josh Sargent' 'Tyler Adams'
 'Brendan Hines-Ike' 'Seyi Adekoya' 'Christian Cappis' 'Ian Harkes'
 'Erik Palmer-Brown' 'Matt Polster' 'Giovanni Reyna' 'Sergino Dest'
 'Yosef Samuel' 'Chris Durkin' 'Dillon Powers' 'Chris Richards'
 'Owen Otasowie' 'Reggie Cannon' 'Luca de la Torre' 'Ethan Horvath'
 'Cameron Carter-Vickers' 'Sebastian Soto']
Mode(most repeated values): Christian Pulisic
Missing values: 0
```

In [52]:
```
info_cat("last_season")
```

```
last_season
Unique values: [2013 2015 2012 2016 2018 2014 2020 2017 2023 2019 2021 202
2]
Mode(most repeated values): 2023
Missing values: 0
```

In [53]:
```
info_cat("current_club_id")
```

```
current_club_id
Unique values: [   42   114  6992   206  2414    86   167  5724    58   64
1  6646   931
   512  1110    24    18   289    20   105    12   995  1084  1063   148
   533  5818    29  2424   379    65   141  1108    38     5  1283    11
   678    79   903   294   506  3999   738   369  3368   385  2671  1123
   989   601 19657  1519   265   124    16   383   475   873  2282  2503
   940   703   371]
Mode(most repeated values): 5
Missing values: 0
```

In [54]:
```
info_cat("player_code")
```

```
player_code
Unique values: ['steven-cherundolo' 'jermaine-jones' 'eugene-starikov' 'cl
arence-goodson'
 'conor-obrien' 'aron-johannsson' 'michael-parkhurst' 'charlie-davies'
 'sacha-kljestan' 'brad-guzan' 'eric-lichaj' 'clint-dempsey'
 'geoff-cameron' 'danny-williams' 'timothy-chandler' 'fabian-johnson'
 'jozy-altidore' 'maurice-edu' 'terrence-boyd' 'michael-bradley'
 'alejandro-bedoya' 'oguchi-onyewu' 'mike-grella' 'bobby-wood'
 'brad-friedel' 'john-anthony-brooks' 'babajide-ogunbiyi' 'tim-howard'
 'caleb-patterson-sewell' 'juan-agudelo' 'sebastian-lletget'
 'julian-green' 'kenny-saief' 'deandre-yedlin' 'matt-miazga'
 'george-fochive' 'alfredo-morales' 'christian-pulisic' 'jordan'
 'desevio-payne' 'gedion-zelalem' 'a-j-soares' 'lynden-gooch'
 'jerome-kiesewetter' 'perry-kitchen' 'emerson-hyndman' 'keaton-parks'
 'jonathan-spector' 'weston-mckennie' 'caleb-stanko' 'emmanuel-sabbi'
 'jonathan-amon' 'shaq-moore' 'andrija-novakovich' 'joseph-efford'
 'timothy-weah' 'tim-ream' 'josh-sargent' 'tyler-adams'
 'brendan-hines-ike' 'seyi-adekoya' 'christian-cappis' 'ian-harkes'
 'erik-palmer-brown' 'matt-polster' 'giovanni-reyna' 'sergino-dest'
 'yosef-samuel' 'chris-durkin' 'dillon-powers' 'chris-richards'
 'owen-otasowie' 'reggie-cannon' 'luca-de-la-torre' 'ethan-horvath'
 'cameron-carter-vickers' 'sebastian-soto']
Mode(most repeated values): christian-pulisic
Missing values: 0
```

In [55]:
```
info_cat("country_of_birth")
```

```
country_of_birth
Unique values: ['United States' 'Germany' 'UdSSR' 'Colombia' 'Italy' 'Engl
and'
 'Netherlands' 'Ethiopia']
Mode(most repeated values): United States
Missing values: 0
```

In [56]:
```
info_cat("date_of_birth")
```

```
date_of_birth
Unique values: <DatetimeArray>
['1979-02-19 00:00:00', '1981-11-03 00:00:00', '1988-11-17 00:00:00',
 '1982-05-17 00:00:00', '1988-10-20 00:00:00', '1990-11-10 00:00:00',
 '1984-01-24 00:00:00', '1986-06-25 00:00:00', '1985-09-09 00:00:00',
 '1984-09-09 00:00:00', '1983-03-09 00:00:00', '1985-07-11 00:00:00',
 '1989-03-08 00:00:00', '1990-03-29 00:00:00', '1987-12-11 00:00:00',
 '1989-11-06 00:00:00', '1986-04-18 00:00:00', '1991-02-16 00:00:00',
 '1987-07-31 00:00:00', '1987-04-29 00:00:00', '1982-05-13 00:00:00',
 '1987-01-23 00:00:00', '1992-11-15 00:00:00', '1971-05-18 00:00:00',
 '1993-01-28 00:00:00', '1986-11-30 00:00:00', '1979-03-06 00:00:00',
 '1987-05-20 00:00:00', '1992-11-23 00:00:00', '1992-09-03 00:00:00',
 '1995-06-06 00:00:00', '1993-12-17 00:00:00', '1993-07-09 00:00:00',
 '1995-07-19 00:00:00', '1992-03-24 00:00:00', '1990-05-12 00:00:00',
 '1998-09-18 00:00:00', '1996-04-26 00:00:00', '1995-11-30 00:00:00',
 '1997-01-26 00:00:00', '1988-11-28 00:00:00', '1995-12-24 00:00:00',
 '1993-02-09 00:00:00', '1992-02-29 00:00:00', '1996-04-09 00:00:00',
 '1997-08-06 00:00:00', '1986-03-01 00:00:00', '1998-08-28 00:00:00',
 '1993-07-26 00:00:00', '1997-12-24 00:00:00', '1999-04-30 00:00:00',
 '1996-11-02 00:00:00', '1996-09-21 00:00:00', '1996-08-29 00:00:00',
 '2000-02-22 00:00:00', '1987-10-05 00:00:00', '2000-02-20 00:00:00',
 '1999-02-14 00:00:00', '1994-11-30 00:00:00', '1995-12-05 00:00:00',
 '1999-08-13 00:00:00', '1995-03-30 00:00:00', '1997-04-24 00:00:00',
 '1993-06-08 00:00:00', '2002-11-13 00:00:00', '2000-11-03 00:00:00',
 '1997-07-03 00:00:00', '2000-02-08 00:00:00', '1991-02-14 00:00:00',
 '2000-03-28 00:00:00', '2001-01-06 00:00:00', '1998-06-11 00:00:00',
 '1998-05-23 00:00:00', '1995-06-09 00:00:00', '1997-12-31 00:00:00',
 '2000-07-28 00:00:00']
Length: 76, dtype: datetime64[ns]
Mode(most repeated values): 1998-09-18 00:00:00
Missing values: 0
```

In [57]:
```
info_cat("sub_position")
```

```
sub_position
Unique values: ['Right-Back' 'Defensive Midfield' 'Centre-Forward' 'Centre
-Back'
 'Attacking Midfield' 'Goalkeeper' 'Second Striker' 'Left Winger'
 'Central Midfield' 'Left Midfield' 'Right Winger' 'Right Midfield']
Mode(most repeated values): Centre-Forward
Missing values: 0
```

In [58]:
```
info_cat("position")
```

```
position
Unique values: ['Defender' 'Midfield' 'Attack' 'Goalkeeper']
Mode(most repeated values): Attack
Missing values: 0
```

In [59]:
```
info_cat("foot")
```

```
foot
Unique values: ['right' 'both' nan 'left']
Mode(most repeated values): right
Missing values: 25
```

In [60]:
```
info_cat("height_in_cm")
```

```
height_in_cm
Unique values: [172. 184. 175. 193. 177.  nan 178. 185. 180. 190. 182. 18
6. 183. 188.
 189. 194. 191. 173. 176. 170. 192. 171. 195.]
Mode(most repeated values): 185.0
Missing values: 16
```

In [61]:
```
info_cat("market_value_in_eur")
```

```
market_value_in_eur
Unique values: [      nan   400000.   100000.   700000.   500000.   75000
0.   200000.
   300000.  2000000.  1000000.   800000. 25000000.  7000000.   150000.
  4000000. 20000000.  1500000.   900000. 14000000. 12000000.   450000.
  2500000. 10000000.  1200000.  3000000. 13000000.   350000.]
Mode(most repeated values): 25000000.0
Missing values: 378
```

In [62]:
```
info_cat("highest_market_value_in_eur")
```

```
highest_market_value_in_eur
Unique values: [ 4000000.  8000000.   500000.  1850000.   750000.  100000
0.  2200000.
  3000000.  5000000.  2000000. 15000000.  2800000.  7000000.  9000000.
  1750000.  6500000.  2500000.   950000.  4500000.   300000. 10000000.
   400000.  3500000. 60000000.   800000.  1800000. 25000000.   600000.
   700000.  1500000.  1700000. 12000000. 20000000.   250000. 42000000.
 30000000.   150000.  1200000. 13000000.   550000.]
Mode(most repeated values): 4000000.0
Missing values: 0
```

In [63]:
```
info_cat("contract_expiration_date")
```

```
contract_expiration_date
Unique values: <DatetimeArray>
[                'NaT', '2024-12-31 00:00:00', '2023-12-31 00:00:00',
 '2025-06-30 00:00:00', '2024-06-30 00:00:00', '2023-11-30 00:00:00',
 '2026-12-31 00:00:00', '2025-12-31 00:00:00', '2027-06-30 00:00:00',
 '2026-06-30 00:00:00', '2028-06-30 00:00:00', '2026-05-31 00:00:00']
Length: 12, dtype: datetime64[ns]
Mode(most repeated values): 2025-06-30 00:00:00
Missing values: 437
```

In [64]:
```
info_cat("agent_name")
```

```
agent_name
Unique values: ['ARP Sportmarketing' nan 'Prosport' 'CAA Stellar' 'Wasserm
an' 'YMU Group'
 'Unique Sports Group' 'YMU Management Ltd.' 'CMG Sports'
 'Robert Schneider' 'TrueSports GmbH' 'Football Company Srl' 'PRO FC'
 'ROGON' 'Mega Sports' 'acta7' 'Promoesport' 'PROSPORT Management'
 'Gestifute' 'athleteMNGment' 'OmniSports' 'SK Soccer Tours'
 'Avid Sports Group' 'FC Enterprise' 'TOP Agency' 'BS Group - BS Law'
 'in4 sports' 'Octagon' 'Joes Blakborn' 'NVA SEG' 'Field Management' 'SBM'
 'CCC']
Mode(most repeated values): Wasserman
Missing values: 499
```

In [65]:
```
info_cat("home_club_manager_name")
```

```
home_club_manager_name
Unique values: ['Michael Wittwer' 'Mirko Slomka' 'Guus Hiddink' 'Miodrag B
ozovic'
 'Rustem Khuzin' 'Kurban Berdyev' 'Auri Skarbalius' 'Lars Sondergaard'
 'Johnny Ma' 'Ariel Jacobs' 'Peter Rensen' 'Keld Bordinggaard'
 'Kasper Hjulmand' 'Kent Nielsen' 'Glen Riddersholm' 'Colin Todd'
 'Troels Bech' 'Adrie Koster' 'John van den Brom' 'Juan Carlos Garrido'
 'Victor Fernandez' 'Alan Pardew' 'Paul Lambert' 'Nigel Adkins'
 'André Villas-Boas' 'Sir Alex Ferguson' 'Martin Jol' "Martin O'Neill"
 'Tony Pulis' 'David Moyes' 'Huub Stevens' 'Markus Babbel' 'Norbert Meier'
 'Jupp Heynckes' 'Bruno Labbadia' 'Christian Streich' 'Dieter Hecking'
 'Sascha Lewandowski' 'Thomas Tuchel' 'Thorsten Fink' 'Thomas Schaaf'
 'Marco Kurz' 'Markus Weinzierl' 'Michael Wiesinger' 'Jens Keller'
 'Lucien Favre' 'Gertjan Verbeek' 'Bülent Uygun' 'Hikmet Karaman'
 'Temur Shalamberidze' nan 'Valdas Urbonas' 'Arsene Wenger'
 'Frank de Boer' 'Steve McClaren' 'Maurice Steijn' 'Art Langeler'
 'Ton Lokhoff' 'Ronald Koeman' 'Ruud Brood' 'Erwin Koeman' 'Peter Bosz'
 'Marco van Basten' 'Jurgen Streppel' 'Zdena k Zeman' 'Roberto Donadoni'
 'Vladimir Petkovic' 'Cristiano Bergodi' 'Eugenio Corini' 'Delio Rossi'
 'Aurelio Andreazzoli' 'Gian Piero Ventura' 'Rafael Benitez'
 'Steve Clarke' 'Chris Hughton' 'Leonardo Jardim' 'Massimiliano Allegri'
 'Manuel Pellegrini' 'Jesualdo Ferreira' 'Martin Jungsgaard'
 'Johan Groote' 'Nicolai Wael' 'Julio Cobos' 'Harry van den Ham'
 'Jurgen Klopp' 'Gaizka Garitano' 'Fatih Terim' 'Jan Poortvliet'
 'Vincenzo Montella' 'Neil Warnock' 'Andrea Stramaccioni' 'Dick Advocaat'
 'Murat Yakin' 'Alois Schwartz' 'Jos Luhukay' 'Robin Dutt' 'Markus Gisdol'
 'Roger Prinzen' 'Niels Frederiksen' 'Ove Christensen' 'sta le solbakken'
 'Francky Dury' 'Glen De Boeck' 'Frédéric Vanderbiest' 'Peter Maes'
 'Michel Der Zakarian' 'Racmi Garde' 'Fabrizio Ravanelli'
 'Jocelyn Gourvennec' 'Élie Baup' 'Herve Renard' 'Claude Puel'
 'Pascal Dupraz' 'Claudio Ranieri' 'Alain Casanova' 'Gustavo Poyet'
 'Roberto Martínez' 'Sam Allardyce' 'Mark Hughes' 'Steve Bruce'
 'Michael Laudrup' 'Rene Meulensteen' 'Jose Mourinho' 'Anatoliy Davydov'
 'Erwin van de Looi' 'Ron Jans' 'Phillip Cocu' 'Jan Wouters'
 'Nebojsa Gudelj' 'Michel Jansen' 'Anton Janssen' 'Jon Dahl Tomasson'
 'Henk Fraser' 'Slaven Bilic' 'Tolunay Kafkas' 'Francesco Guidolin'
 'Rudi Garcia' 'Georgios Paraschos' 'Jorge Jesus' 'Dennis Haar'
 'Carsten Mikkelsen' 'Laurent Blanc' 'Laurentiu Reghecampf' 'Michel'
 'Billy Davies' 'Eric Hellemons' 'Henk de Jong' 'Rob Alflen'
 'Dwight Lodeweges' 'Alfred Schreuder' 'Marinus Dijkhuizen'
 'Tayfun Korkut' 'Pal Dardai' 'Viktor Skrypnyk' 'Rene Girard'
 'Claude Makelele' 'Sylvain Ripoll' 'Dominique Arribage' 'Brendan Rodgers'
 'Hein Vanhaezebrouck' 'Erik Assink' 'Yves Vanderhaeghe' 'Felice Mazzu'
 'Rachid Chihab' 'Aleksandar Jankovic' 'Quique Sanchez Flores'
 'Alexander Zorniger' 'Ove Pedersen' 'Jakob Michelsen' 'Andre Schubert'
 'Dirk Schuster' 'Ralph Hasenhuttl' 'Roger Schmidt' 'Andre Breitenreiter'
 'Jurgen Kramny' 'Willy Sagnol' 'Christophe Galtier' 'Olivier Gevaert'
 'Harm van Veldhoven' 'Vyacheslav Groznyi' 'Oleksandr Babych'
 'Ghislain Printant' 'Thierry Laurey' 'Hubert Fournier' 'Zoran Mamic'
 'Rene Weiler' 'Olafur Helgi Kristjansson' 'Bo Henriksen' 'Jess Thorup'
 'Lars Lungi Sorensen' 'Pep Guardiola' 'Pavel Dotchev' 'Heiko Herrlich'
 'John Stegeman' 'Giannis Anastasiou' 'Mitchell van der Gaag'
 'Carlo Ancelotti' 'Markus Kauczinski' 'Martin Schmidt'
 'Julian Nagelsmann' 'Maik Walpurgis' 'Torsten Frings' 'Niko Kovac'
 'Peter Stoger' 'Andries Jonker' 'Robbie Neilson' 'Drazen Besek'
 'Lee Clark' 'Graeme Murty' 'Mark McGhee' 'Ian Cathro' 'Lee McCulloch'
 'Pedro Caixinha' 'Jacky Mathijssen' 'Rudi Cossey' 'Yannick Ferrera'
 'Runar Kristinsson' 'Morten Rasmussen' 'Rico Schmitt' 'Zinedine Zidane'
 'Henrik Lehm' 'Janos Radoki' 'Neil Harris' 'Ismail Atalan' 'Neil Lennon'
 'Joe Enochs' 'Dirk Heyne' 'Florian Schnorrenberg' 'Rene Rydlewicz'
 'Manuel Baum' 'Domenico Tedesco' 'Stefan Ruthenbeck' 'Florian Kohfeldt'
```

```
 'Alexander Nouri' 'Sandro Schwarz' 'Eddie Howe' 'Sean Dyche'
 'Christian Lanstrup' 'Stale Solbakken' 'Thomas Thomasberg'
 'Edward Sturing' 'Reinier Robbemond' 'Juan Claudio'
 'Gian Piero Gasperini' 'Mark Robins' 'Phil Parkinson' 'David Wagner'
 'Antoine Kombouare' 'Sabri Lamouchi' 'Patrick Vieira'
 'Christophe Pellissier' 'Bruno Genesio' 'Julien Stephan' 'Oliver Barth'
 'Slavisa Jokanovic' 'Unai Emery' 'Maurizio Sarri' 'Roy Hodgson'
 'Javi Gracia' 'Rene Eijer' 'Friedhelm Funkel' 'Michael Kollner '
 'Ralf Rangnick' 'Adolfo Sormani' 'Allan Kuhn' 'Claus Aagaard'
 'Jens Berthel Askou' 'David Nielsen' 'John Lammers' 'Jacob Friis'
 'Tommy Wright' 'Oran Kearney' 'Steven Gerrard' 'Zoran Mirkovic'
 'Holger Bachthaler' 'Ivan Leko' 'Yuri Semin' 'Diego Simeone'
 'Aleksandr Khatskevich' 'Nenad Bjelica' 'Morten Karlsen'
 'Mauricio Pochettino' 'Giovanni van Bronckhorst' 'Mark van Bommel'
 'Danny Buijs' 'Erik ten Hag' 'Frank Wormuth' 'Key Riebau'
 'Olivier Dall Oglio' 'Vincent Hognon' 'David Guion' 'Marco Rose'
 'Adi Hutter' 'Uwe Seeler' 'Jurgen Klinsmann' 'Hansi Flick'
 'Steffen Baumgart' 'Oliver Glasner' 'Kenneth Andersen'
 'Flemming Pedersen' 'Daniel Farke' 'Frank Lampard' 'Chris Wilder'
 'Dean Smith' 'Graham Potter' 'Uwe Neuhaus' 'Christian Flindt Bjerg'
 'Albert Celades' 'Nicky Hayen' 'Rudi Vata' 'Dennis van Wijk'
 'Milos Kostic' 'Stephen Robinson' 'Mikel Arteta' 'Torsten Lieberknecht'
 'Philippe Clement' 'Hernan Losada' 'Wouter Vrancken' 'Dirk Kunert']
Mode(most repeated values): Gertjan Verbeek
Missing values: 727
```

In [66]:
```
info_cat("away_club_manager_name")
```

```
away_club_manager_name
Unique values: ['Mirko Slomka' 'Huub Stevens' 'Miodrag Bozovic' 'Vladimir
Bibikov'
 'Jess Thorup' 'Troels Bech' 'Peter Rensen' 'Kasper Hjulmand'
 'Lars Sondergaard' 'Glen Riddersholm' 'Colin Todd' 'Auri Skarbalius'
 'Ariel Jacobs' 'Keld Bordinggaard' 'Johnny Ma' 'Kent Nielsen'
 'Viggo Jensen' 'John van den Brom' 'Philippe Clement' 'Luka Peruzovic'
 'Mircea Rednic' 'Paul Lambert' 'Michael Laudrup' 'Mark Hughes'
 'andre villas boas' 'Roberto Di Matteo' 'Roberto Martínez' 'Martin Jol'
 'Alan Pardew' 'Brendan Rodgers' 'Tony Pulis' 'Paolo Di Canio'
 'David Moyes' 'Sir Alex Ferguson' 'Roberto Mancini' 'Mauricio Pochettino'
 'Markus Weinzierl' 'Dieter Hecking' 'Markus Babbel' 'Jupp Heynckes'
 'Armin Veh' 'Jurgen Klopp' 'Thomas Schaaf' 'Sascha Lewandowski'
 'Lorenz Gunther Kostner' 'Frank Kramer' 'Bruno Labbadia' 'Thomas Tuchel'
 'Thorsten Fink' 'Mike BUskens' 'Norbert Meier' 'Michael Wiesinger'
 'Jens Keller' 'Marco Kurz' 'Markus Gisdol' 'Temur Shalamberidze'
 'Gert Heerkes' 'Hikmet Karaman' 'Yilmaz Vural' 'Hamza Hamzaoglu' nan
 'Brian McDermott' 'Gertjan Verbeek' 'Peter Bosz' 'Marco van Basten'
 'Ruud Brood' 'Erwin Koeman' 'Fred Rutten' 'Robert Maaskant'
 'Frank de Boer' 'Jan Wouters' 'Dick Advocaat' 'Art Langeler'
 'Rolando Maran' 'Stefano Colantuono' 'Zdena  Zeman' 'Aurelio Andreazzoli'
 'Davide Ballardini' 'Pampos Christodoulou' 'Guus Hiddink' 'Orest Lenczyk'
 'Arsene Wenger' 'Mircea Lucescu' 'Vladimir Petkovic' 'Age Hareide'
 'juan ramon rocha' 'Manuel Pellegrini' 'Racmi Garde' 'Mark Robins'
 'Ian Holloway' 'Jos Luhukay' 'Robin Dutt' 'Thomas Schneider'
 'Pep Guardiola' 'Roger Prinzen' 'Torsten Lieberknecht' 'Tayfun Korkut'
 'Adrie Bogers' 'Ove Christensen' 'Niels Frederiksen' 'Victor Fernandez'
 'Hein Vanhaezebrouck' 'Harm van Veldhoven' 'Bob Peeters'
 'Ronny Van Geneugden' 'Guy Luzon' 'Cedomir Janevski' 'Laurent Blanc'
 'Eric Ly' 'Michel Der Zakarian' 'Alain Casanova' 'Christian Gourcuff'
 'Hubert Fournier' 'Francis Gillot' 'Jocelyn Gourvennec'
 'Christophe Galtier' 'Chris Hughton' 'Gustavo Poyet' 'Jose Mourinho'
 'Rene Meulensteen' 'Leonid Slutski' 'Dennis Haar' 'Foeke Booy'
 'Phillip Cocu' 'Dwight Lodeweges' 'Michel Jansen' 'Nebojsa Gudelj'
 'Ron Jans' 'Jan de Jonge' 'Erwin van de Looi' 'Anton Janssen'
 'Samet Aybaba' 'Ertugrul Saglam' 'Slaven Bilic' 'Ertugrul Secme'
 'Ersun Yanal' 'Mehmet Ozdilek' 'Peter Maes' 'Rudi Garcia'
 'Luigi De Canio' 'Georgios Paraschos' 'Viktor Kumykov' 'Arik Benado'
 'Jaroslav Silhavy' 'Sam Allardyce' 'Gadzhi Gadzhiev' 'John Stegeman'
 'Henk de Jong' 'AndrE Breitenreiter' 'Lucien Favre' 'Pa Dardai'
 'Roger Schmidt' 'Antoine Kombouare' 'Rolland Courbis' 'Patrice Garande'
 'Marcelino' 'Rachid Chihab' 'Besnik Hasi' 'Kevin Wilkin'
 'Frederic Vanderbiest' 'Olivier Guillou' 'Quique Sanchez Flores'
 'Louis van Gaal' 'Viktor Skrypnyk' 'Jonas Dal' 'Ralph Hasenhuttl'
 'andrea schubert' 'Martin Schmidt' 'Niko Kovac' 'Daniel Stendel'
 'Olivier Gevaert' 'Oleksandr Babych' 'Vasyl Sachko' 'Dominique Arribage'
 'Anatoliy Chantsev' 'Volodymyr Mazyar' 'Roman Sanzhar' 'Willy Sagnol'
 'Thierry Laurey' 'Massimiliano Allegri' 'Nuno Espirito Santo'
 'Unai Emery' 'Volodymyr Sharan' 'felice mazzu' 'Francky Dury'
 'Bo Henriksen' 'David Nielsen' 'Alexander Zorniger' 'Ronald Koeman'
 'Markus Kauczinski' 'Ernest Faber' 'Henk Fraser' 'Dirk Schuster'
 'Carlo Ancelotti' 'Maik Walpurgis' 'Christian Streich' 'Peter Stoger'
 'Torsten Frings' 'Manuel Baum' 'Julian Nagelsmann' 'Richie Foran'
 'Robbie Neilson' 'Tommy Wright' 'Mark Warburton' 'Ian Cathro'
 'Pedro Caixinha' 'Martin Canning' 'Georges Leekens' 'Karim Belhocine'
 'Bart Wilmssen' 'Aykut Kocaman' 'Hannes Wolf' 'jose peseiro'
 'Dennis van den IJssel' 'Rui Vitoria' 'Paulo Sousa' 'Eddie Howe'
 'Jim Duffy' 'Albert Stuivenberg' 'Andries Jonker' 'Domenico Tedesco'
 'Alfons Groenendijk' 'Bernd Hollerbach' 'Florian Kohfeldt'
 'Heiko Herrlich' 'Alexander Nouri' 'Christian Titz' 'Sandro Schwarz'
 'David Wagner' 'Marco Silva' 'Olafur Kristjansson' 'Thomas Thomasberg'
```

```
 'Marino Pusic' 'Edward Sturing' 'Cuco Ziganda' 'Sven Vermant'
 'Gian Piero Gasperini' 'Paul Cook' 'Steve Cotterill' 'Fabien Mercadal'
 'Eric Bedouet' 'Sabri Lamouchi' 'Julien Stephan' 'Leonardo Jardim'
 'Bernard Blaquart' 'Jean-Louis Gasset' 'Maurizio Sarri' 'Rafael Benatez'
 'Claude Puel' 'Claudio Ranieri' 'Scott Parker' 'Mark van Bommel'
 'Jean-Paul de Jong' 'Ralf Rangnick' 'Friedhelm Funkel' 'Michael Kollner'
 'Adi HUtter' 'Adrie Koster' 'Allan Kuhn' 'Jakob Michelsen'
 'Jens Berthel Askou' 'Adolfo Sormani' 'Claus Aagaard ' 'Stale Solbakken'
 'Morten Wieghorst' 'Kenneth Andersen' 'Danny Buijs' 'Neil Lennon'
 'Steve Clarke' 'Gary Holt' 'Oran Kearney' 'Urs Fischer' 'Diego Simeone'
 'Stanimir Stoilov' 'Marc Brys' 'Rene Eijer' 'Ian Murray'
 'Adrie Poldervaart' 'Mitchell van der Gaag' 'Dick Lukkien'
 'Maurice Steijn' 'Giovanni van Bronckhorst' 'Frank Lampard'
 'Patrick Vieira' 'Derek McInnes' 'Marco Rose' 'Ante Covic'
 'Achim Beierlorzer' 'Alfred Schreuder' 'Oliver Glasner' 'Uwe Seeler'
 'Marcel Rapp' 'Hansi Flick' 'Flemming Pedersen' 'Lars Olsen'
 'Sjors Ultee' 'Erik ten Hag' 'Jaap Stam' 'Daniel Farke' 'Dean Smith'
 'Roy Hodgson' 'Simone Inzaghi' 'Milos Kostic' 'Bea San Jose'
 'Bernd Storck' 'Luis Castro' 'Micky Mellon' 'Philippe Montanier'
 'Nicky Hayen' 'Kevin Muscat' 'Frank Wormuth']
Mode(most repeated values): Gertjan Verbeek
Missing values: 727
```

In [68]:
```
#filling missing values
```

In [69]:
```
df_event["stadium"].isnull().sum()
```

Out[69]: 723

In [ ]:
```

```

In [70]:
```
df_event["stadium"].fillna("AFAS Stadion", inplace=True)
```

In [71]:
```
df_event["stadium"].isnull().sum()
```

Out[71]: 0

In [72]:
```
df_event["description"].isnull().sum()
```

Out[72]: 933

In [73]:
```
df_event["description"].fillna(df_event["description"].mode()[0], inplace=Tr
```

In [74]:
```
df_event["description"].isnull().sum()
```

Out[74]: 0

In [75]:
```
df_event["referee"].isnull().sum()
```

Out[75]: 725

In [76]:
```
df_event["referee"].fillna("Felix Zwayer",inplace = True)
```

In [77]:
```
df_event["referee"].isnull().sum()
```

Out[77]: 0

In [78]:
```
df_event["season"].isnull().sum()
```

Out[78]: 723

In [79]:
```
df_event["season"].fillna("2012.0", inplace = True)
```

In [80]:
```
df_event["season"].isnull().sum()
```

Out[80]: 0

In [81]:
```
df_event["round"].fillna(df_event["round"].mode()[0], inplace = True)
```

In [82]:
```
df_event["round"].isnull().sum()
```

Out[82]: 0

In [83]:
```
df_event["home_club_goals"].isnull().sum()
```

Out[83]: 723

In [84]:
```
df_event["home_club_goals"].fillna(df_event["home_club_goals"].mode()[0], ir
```

In [85]:
```
df_event["home_club_goals"].mean()
```

Out[85]: 1.5186587344510547

In [86]:
```
df_event["home_club_goals"].isnull().sum()
```

Out[86]: 0

In [ ]:
```

```

In [87]:
```
df_event["away_club_goals"].fillna(df_event["away_club_goals"].mode()[0], ir
```

In [88]:
```
df_event["away_club_goals"].isnull().sum()
```

Out[88]: 0

In [89]:
```
df_event["home_club_position"].fillna(df_event["home_club_position"].mode()[
```

In [90]:
```
df_event["home_club_position"].isnull().sum()
```

Out[90]: 0

In [ ]:
```

```

In [91]:
```
df_event["away_club_position"].fillna(df_event["away_club_position"].mode()[
```

In [92]:
```
df_event["away_club_position"].isnull().sum()
```

Out[92]: 0

In [93]:
```
df_event["aggregate"].fillna(df_event["aggregate"].mode()[0], inplace = True
```

In [94]:
```
df_event["aggregate"].isnull().sum()
```

Out[94]: 0

In [95]:
```
df_event["competition_type"].fillna(df_event["competition_type"].mode()[0],
```

In [96]:
```
df_event["competition_type"].isnull().sum()
```

Out[96]: 0

In [97]:
```
df_event["foot"].fillna(df_event["foot"].mode()[0], inplace = True)
```

In [98]:
```
df_event["foot"].isnull().sum()
```

Out[98]: 0

In [99]:
```
df_event["height_in_cm"].fillna(df_event["height_in_cm"].mean(), inplace = T
```

In [100]:
```
df_event["height_in_cm"].isnull().sum()
```

Out[100]: 0

In [101]:
```
df_event["market_value_in_eur"].fillna(df_event["market_value_in_eur"].mean(
```

In [102]:
```
df_event["market_value_in_eur"].isnull().sum()
```

Out[102]: 0

In [103]:
```
df_event["contract_expiration_date"].fillna(df_event["contract_expiration_da
```

In [104]:
```
df_event["contract_expiration_date"].isnull().sum()
```

Out[104]: 0

In [105]:
```
df_event["agent_name"].fillna(df_event["agent_name"].mode()[0], inplace = Tr
```

In [106]:
```
df_event["agent_name"].isnull().sum()
```

Out[106]: 0

In [107]:
```
df_event["home_club_name"].isnull().sum()
```

Out[107]: 777

In [108]:
```
df_event["home_club_name"].fillna(df_event["home_club_name"].mode()[0], inpl
```

In [109]:
```
df_event["home_club_name"].isnull().sum()
```

Out[109]: 0

In [110]:
```
df_event["away_club_name"].isnull().sum()
```

Out[110]: 740

In [111]:
```
df_event["away_club_name"].fillna(df_event["away_club_name"].mode()[0], inp
```

In [112]:
```
df_event["home_club_manager_name"].isnull().sum()
```

Out[112]: 727

In [113]:
```
df_event["home_club_manager_name"].fillna(df_event["home_club_manager_name"
```

In [114]:
```
df_event["away_club_manager_name"].isnull().sum()
```

Out[114]: 727

In [115]:
```
df_event["away_club_manager_name"].fillna(df_event["away_club_manager_name"
```

In [ ]:
```

```

In [116]:
```
df_event.isnull().sum()
```

Out[116]:
```
game_event_id                  0
date_x                         0
game_id                        0
minute                         0
type                           0
player_id                      0
description                    0
player_in_id                 694
player_assist_id            1709
competition_id               723
season                         0
round                          0
date_y                       723
home_club_goals                0
away_club_goals                0
home_club_position             0
away_club_position             0
home_club_manager_name         0
away_club_manager_name         0
stadium                        0
attendance                   769
referee                        0
home_club_formation         1849
away_club_formation         1849
home_club_name                 0
away_club_name                 0
aggregate                      0
competition_type               0
name                           0
last_season                    0
current_club_id                0
player_code                    0
country_of_birth               0
date_of_birth                  0
sub_position                   0
position                       0
foot                           0
height_in_cm                   0
market_value_in_eur            0
highest_market_value_in_eur    0
contract_expiration_date       0
agent_name                     0
dtype: int64
```

In [ ]:
```

```

In [117]:
```
#for numerical data
```

In [118]:
```
def info_num(col):                    #numerical
    print(col)
    print("Mean:",df_event[col].mean())
    print("Median:",df_event[col].median())
    print("Mode", df_event[col].mode()[0])
    print("Standard deviation:",df_event[col].std())
    print("Minimum value:",df_event[col].min())
    print("Maximum value:",df_event[col].max())
    print("Missing value:",df_event[col].isnull().sum())
```

In [ ]:
```
```

In [119]:
```
info_num("game_id")
```

```
game_id
Mean: 3023615.0070308275
Median: 3058692.0
Mode 2224542
Standard deviation: 592952.732777134
Minimum value: 2221641
Maximum value: 4194154
Missing value: 0
```

In [120]:
```
info_num("date_x")
```

```
date_x
Mean: 2018-03-30 10:57:18.399134720
Median: 2018-10-06 00:00:00
Mode 2021-04-10 00:00:00
Standard deviation: 1327 days 23:39:38.261191088
Minimum value: 2012-07-14 00:00:00
Maximum value: 2023-11-12 00:00:00
Missing value: 0
```

In [121]:
```
info_num("minute")
```

```
minute
Mean: 63.75283937263386
Median: 69.0
Mode 90
Standard deviation: 21.6402608938712
Minimum value: -1
Maximum value: 110
Missing value: 0
```

In [122]:
```
info_num("player_in_id")
```

```
player_in_id
Mean: 239512.7038961039
Median: 187492.0
Mode 117432.0
Standard deviation: 188956.03455020505
Minimum value: 411.0
Maximum value: 1028162.0
Missing value: 694
```

In [123]:
```
info_num("date_y")
```

```
date_y
Mean: 2016-01-23 04:23:26.749555968
Median: 2015-12-04 12:00:00
Mode 2012-10-07 00:00:00
Standard deviation: 963 days 06:42:20.779492032
Minimum value: 2012-07-14 00:00:00
Maximum value: 2020-09-26 00:00:00
Missing value: 723
```

In [124]:
```
info_num("attendance")
```

```
attendance
Mean: 27686.618518518517
Median: 23696.0
Mode 81360.0
Standard deviation: 20317.982483026317
Minimum value: 300.0
Maximum value: 81365.0
Missing value: 769
```

In [126]:
```
df_event["attendance"].fillna(df_event["attendance"].mean(), inplace = True)
```

In [127]:
```
df_event["attendance"].isnull().sum()
```

Out[127]: 0

In [128]:
```
df_event["date_y"].fillna(df_event["date_y"].mode()[0], inplace = True)
```

In [129]:
```
df_event["date_y"].isnull().sum()
```

Out[129]: 0

In [122]:
```
df_event.isnull().sum()
```

Out[122]:
```
game_event_id                  0
date_x                         0
game_id                        0
minute                         0
type                           0
player_id                      0
description                    0
player_in_id                 694
player_assist_id            1709
competition_id               723
season                         0
round                          0
date_y                         0
home_club_goals                0
away_club_goals                0
home_club_position             0
away_club_position             0
home_club_manager_name         0
away_club_manager_name         0
stadium                        0
attendance                     0
referee                        0
home_club_formation         1849
away_club_formation         1849
home_club_name                 0
away_club_name                 0
aggregate                      0
competition_type               0
name                           0
last_season                    0
current_club_id                0
player_code                    0
country_of_birth               0
date_of_birth                  0
sub_position                   0
position                       0
foot                           0
height_in_cm                   0
market_value_in_eur            0
highest_market_value_in_eur    0
contract_expiration_date       0
agent_name                     0
dtype: int64
```

In [ ]:
```
#removing duplicate
```

In [130]:
```
columns_to_exclude = ["player_id", "game_id", "player_assist_id", "competiti
df_event = df_event.drop_duplicates(subset = [col for col in df_event.columr
```

In [ ]:

In [ ]:

In [ ]:

In [131]:
```
df_event
```

Out[131]:

	game_event_id	date_x	game_id	minute	type	player_id
0	c6a3c088ed8a38d4ce074dd73b20d3da	2012-08-19	2221641	62	Substitutions	1335
1	02d605a5c2dc4f9a6721daa583fa5405	2012-08-26	2222536	54	Cards	1321
2	b56c2e2e087cddb3cfe9e3d340975df9	2012-11-18	2222707	79	Substitutions	104203
3	4a15d1fff4f476f48bb60092c61641d5	2012-11-23	2222721	72	Substitutions	104203
4	daa97877f7edf2fda885b411d7197921	2013-05-17	2222782	63	Goals	104203
...	...	...	...	...	...	...
1842	4d7035745771b8e12e5ad708edef421e	2023-	4171200	90	Substitutions	315770

In [132]:
```
#Coverting the dataset into csv file
```

In [179]:
```
df_event.to_csv("df_event_cleaned.csv")
```

In [126]:
```
df_event.isnull().sum()
```

Out[126]:
```
game_event_id                0
date_x                       0
game_id                      0
minute                       0
type                         0
player_id                    0
description                  0
player_in_id               693
player_assist_id          1567
competition_id             581
season                       0
round                        0
date_y                       0
home_club_goals              0
away_club_goals              0
home_club_position           0
away_club_position           0
home_club_manager_name       0
away_club_manager_name       0
```

In [ ]:

For df_appear

In [133]:
```
df_appear
```

Out[133]:

	appearance_id	game_id	player_id	date_x	player_name	competition_id_x	yellow_car
0	2224728_119169	2224728	119169	2012-07-13	Aron Johannsson	DK1	
1	2224732_161244	2224732	161244	2012-07-14	Conor O'Brien	DK1	
2	2224729_39467	2224729	39467	2012-07-15	Clarence Goodson	DK1	
3	2232104_119169	2232104	119169	2012-07-19	Aron Johannsson	ELQ	
4	2219794_39475	2219794	39475	2012-07-22	Sacha Kljestan	BESC	
...	...	...	...	...	...	...	
3563	3415291_537467	3415291	537467	2020-09-26	Joseph Efford	BE1	
3564	3415296_367423	3415296	367423	2020-09-26	Chris Durkin	BE1	
3565	3431983_478940	3431983	478940	2020-09-26	Reggie Cannon	PO1	
3566	3450575_361104	3450575	361104	2020-09-26	Sergino Dest	NL1	
3567	3412904_124732	3412904	124732	2020-09-27	John Anthony Brooks	L1	

3568 rows × 44 columns

In [134]:
```
df_appear.isnull().sum()
```

Out[134]:
```
appearance_id                  0
game_id                        0
player_id                      0
date_x                         0
player_name                    0
competition_id_x               0
yellow_cards                   0
red_cards                      0
goals                          0
assists                        0
minutes_played                 0
competition_id_y               0
season                         0
round                          0
date_y                         0
home_club_goals                0
away_club_goals                0
home_club_position           503
away_club_position           503
home_club_manager_name         5
away_club_manager_name         5
stadium                        0
attendance                   133
referee                        3
home_club_formation         3568
away_club_formation         3568
home_club_name               113
away_club_name                65
aggregate                      0
competition_type               0
name                           0
last_season                    0
current_club_id                0
player_code                    0
country_of_birth               0
date_of_birth                  0
sub_position                   0
position                       0
foot                          90
height_in_cm                  51
market_value_in_eur         1220
highest_market_value_in_eur    0
contract_expiration_date    1348
agent_name                  1021
dtype: int64
```

In [137]:
```
#Separating the data into categorical and continuous
```

In [138]:
```
def sep_data_types(df_appear):
    categorical = []
    continuous = []
    for column in df_appear.columns:
        if df_appear[column].nunique() < 100:
            categorical.append(column)
        else:
            continuous.append(column)

    return categorical, continuous

categorical, continuous = sep_data_types(df_appear)

#Tabulate is a package used to print the list, dict or any data sets in a pr

from tabulate import tabulate
table = [categorical, continuous]
print(tabulate({"Categorical":categorical, "Continuous":continuous}, headers
```

```
categorical                 continuous
--------------------------  --------------------
player_id                   appearance_id
player_name                 game_id
competition_id_x            date_x
yellow_cards                date_y
red_cards                   home_club_manager_name
goals                       away_club_manager_name
assists                     stadium
minutes_played              attendance
competition_id_y            referee
season                      home_club_name
round                       away_club_name
home_club_goals
away_club_goals
home_club_position
away_club_position
home_club_formation
away_club_formation
aggregate
competition_type
name
last_season
current_club_id
player_code
country_of_birth
date_of_birth
sub_position
position
foot
height_in_cm
market_value_in_eur
highest_market_value_in_eur
contract_expiration_date
agent_name
```

In [139]:
```
def information_cat(col):           #categorical
    print(col)
    print("Unique values:", df_appear[col].unique())
    print("Mode(most repeated values):", df_appear[col].mode()[0])
    print("Missing values:",df_appear[col].isnull().sum())
```

In [140]:
```
information_cat("agent_name")
```

```
agent_name
Unique values: ['CAA Stellar' nan 'Wasserman' 'ARP Sportmarketing' 'Robert
Schneider'
 'Unique Sports Group' 'CMG Sports' 'Football Company Srl' 'YMU Group'
 'YMU Management Ltd.' 'Prosport' 'ROGON' 'acta7' 'Mega Sports'
 'athleteMNGment' 'BR Group Management' 'Gestifute' 'PROSPORT Management'
 'AKA Global GmbH' 'Promoesport' 'fair-sport GmbH' 'PRO FC' 'OmniSports'
 'TrueSports GmbH' 'SBM' 'Avid Sports Group' 'FC Enterprise' 'Octagon'
 'BS Group - BS Law' 'CAA Base Ltd' 'in4 sports' 'CCC' 'SK Soccer Tours'
 'Joes Blakborn' 'PG 120 Sport Agency' 'NVA SEG' 'GROW' 'TOP Agency'
 'Field Management']
Mode(most repeated values): Wasserman
Missing values: 1021
```

In [141]:
```
information_cat("contract_expiration_date")
```

```
contract_expiration_date
Unique values: <DatetimeArray>
['2024-12-31 00:00:00',                 'NaT', '2025-06-30 00:00:00',
 '2023-12-31 00:00:00', '2024-06-30 00:00:00', '2026-12-31 00:00:00',
 '2023-11-30 00:00:00', '2025-12-31 00:00:00', '2027-06-30 00:00:00',
 '2026-05-31 00:00:00', '2028-06-30 00:00:00', '2026-06-30 00:00:00']
Length: 12, dtype: datetime64[ns]
Mode(most repeated values): 2024-06-30 00:00:00
Missing values: 1348
```

In [142]:
```
information_cat("market_value_in_eur")
```

```
market_value_in_eur
Unique values: [  400000.       nan   500000.   700000.   200000.   10000
0.  2000000.
   300000.  1000000.   150000.  7000000.   800000. 25000000. 13000000.
   750000. 20000000.  1500000.   250000.  4000000. 14000000.  2500000.
   900000. 12000000.   350000. 10000000.  3500000.  1200000.   450000.
  3000000.]
Mode(most repeated values): 2000000.0
Missing values: 1220
```

In [143]:
```
information_cat("height_in_cm")
```

```
height_in_cm
Unique values: [184. 177. 193. 185. 178.  nan 172. 186. 190. 191. 183. 18
0. 189. 182.
 175. 179. 194. 173. 176. 170. 188. 195. 162. 192. 171.]
Mode(most repeated values): 183.0
Missing values: 51
```

In [144]:
```
information_cat("foot")
```

```
foot
Unique values: ['right' 'both' nan 'left']
Mode(most repeated values): right
Missing values: 90
```

In [145]:
```
information_cat("away_club_name")
```

```
away_club_name
Unique values: ['Aalborg BK' 'Randers Fodbold Club' 'Odense Boldklub' na
n
 'KSC Lokeren (- 2020)' 'BRA NDBYERNES' 'Football Club Ka Benhavn'
 'Esbjerg fB' 'Aarhus Gymnastik Forening' 'Sonderjyske Fc' 'Silkeborg I
F'
 'Hannover 96' 'Fodbold Club Nords  Aalnd' 'Beerschot AC' 'AC Horsens'
 'Royal Sporting Club Anderlecht' 'Alkmaar Zaanstreek'
 'RAEC Mons (- 2015)' 'Tottenham Hotspur Football Club' '1.FC Nuremberg'
 'Heracles Almelo' ' Vitoria Setubal FC' 'Manchester United Football Clu
b'
 'tsg 1899 hoffenheim football spielbetriebs Gmbh' 'Everton Football Clu
b'
 'West Bromwich Albion' 'FC Schalke 04' 'Arsenal Football Club'
 'Sport Lisboa e Benfica' 'Sportclub Heerenveen' 'Catania FC'
 'Anzhi Makhachkala ( -2022)' 'Norwich City' 'Stoke City'
 'eintracht frankfurt Football ag' 'Borussia Dortmund' 'FC Augsburg 190
7'
 'Koninklijke Racing Club Genk' 'Aston Villa Football Club'
```

In [146]:
```
information_cat("home_club_name")
```

```
home_club_name
Unique values: ['Aarhus Gymnastik Forening' 'Sonderjyske Fc' 'Brondby I
F'
 'Royal Sporting Club Anderlecht' 'Aalborg BK' 'Odense Boldklub' nan
 'Esbjerg fB' 'Randers Fodbold Club' 'Football Club Nords'
 'Fodbold Club Midtjylland' 'Hannover 96' 'F.C. Copenhagen'
 'Cercle Brugge Koninklijke Sportvereniging' 'AFC Ajax Amsterdam'
 'Silkeborg Idrra Sforening' 'Newcastle United Football Club'
 'Alkmaar Zaanstreek' 'CD Nacional' 'Everton Football Club'
 'Anzhi Makhachkala ( -2022)' 'Borussia Verein für Leibesübung 1900 e.
V.'
 'Hamburger SV' 'Oud-Heverlee Leuven' 'Aston Villa Football Club'
 'Tottenham Hotspur Football Club' 'Stoke City' 'fc vitoria setubal'
 'Associazione Sportiva Roma' 'AC Horsens' 'West Bromwich Albion'
 'Wigan Athletic' 'TSG 1899 Hoffenheim Football-Spielbetriebs GmbH'
 '1.FC Nuremberg' 'FC Schalke 04' 'verein fur leibesubungen'
 'Eindhovense Voetbalvereniging Philips Sport Vereniging'
 'Lierse SK (- 2018)' 'spvgg greuther furth games' 'Reading FC'
 'Sport-Club Freiburg' 'Olympiakos Syndesmos Filathlon Peiraios'
```

In [147]:
```
information_cat("referee")
```

```
referee
Unique values: ['Michael Svendsen' 'Claus Bo Larsen' 'Lars Christofferse
n'
 'Boako Jovanetic' 'Laurent Colemonts' 'Mads-Kristoffer Kristoffersen'
 'Michael Tykgaard' 'Jens Maae' 'Stephan Studer' 'Michael Johansen'
 'Jakob Kehlet' 'Kristinn Jakobsson' 'Tamas Bognair' 'Kenn Hansen'
 'Sacbastien Delferia re' 'Henning Jensen' 'Alon Yefet' 'Anar Salmanov'
 'Peter Rasmussen' 'Henrik Kragh' 'Luc Wouters' 'Kevin Blom'
 'Joeri van de Velde' 'Martin Atkinson' 'Benjamin Cortus' 'Robert Kempte
r'
 'Serdar Gaza' 'Jorge Ferreira' 'Andre Marriner' 'Milorad Mazic'
 'Michalis Koukoulakis' 'Paolo Tagliavento' 'Anders Poulsen'
 'Peter Sippel' 'Marco Fritz' 'Serge Gumienny' 'Michael Oliver'
 'Mike Dean' 'Dr. Felix Brych' 'Lee Mason' 'Jorge Sousa' 'Danny Makkeli
e'
 'Andrea De Marco' 'Marcin Borski' 'Svein Oddvar Moen' 'Sergey Karasev'
 'Mark Halsey' 'Jonathan Moss' 'Michael Weiner' 'Tobias Welz'
 'Daniel Siebert' 'Tim Pots' 'Lee Probert' 'Wolfgang Stark' 'Bas Nijhui
s'
```

In [148]:
```
information_cat("home_club_manager_name")
```

```
home_club_manager_name
Unique values: ['Peter Rensen' 'Lars Sondergaard' 'Auri Skarbalius' 'Joh
n van den Brom'
 'Kent Nielsen' 'Troels Bech' 'Temur Shalamberidze' 'Jess Thorup'
 'Colin Todd' 'Kasper Hjulmand' 'Liam Buckley' 'Glen Riddersholm'
 'Valdas Urbonas' 'Mirko Slomka' 'Ariel Jacobs' 'Bob Peeters'
 'Frank de Boer' 'Keld Bordinggaard' 'Alan Pardew' 'Michael Wittwer'
 'Andre Breitenreiter' 'Gertjan Verbeek' 'Pedro Caixinha' 'David Moyes'
 'Pampos Christodoulou' 'Guus Hiddink' 'Orest Lenczyk' 'Lucien Favre'
 'Thorsten Fink' 'Ronny Van Geneugden' 'Paul Lambert' 'André Villas-Boa
s'
 'Tony Pulis' 'jose mota' 'Zdena k Zeman' 'Johnny Ma' 'Steve Clarke'
 'Roberto Martínez' 'Markus Babbel' 'Dieter Hecking' 'Huub Stevens'
 'Felix Magath' 'Dick Advocaat' 'Chris Janssens' 'Mike Buskens'
 'Brian McDermott' 'Christian Streich' 'Leonardo Jardim'
 'Massimiliano Allegri' 'Mircea Lucescu' 'Steve McClaren' 'Francky Dury'
 'Roberto Di Matteo' 'Michael Laudrup' 'Nigel Adkins' 'John Karelse'
 'Martin Jungsgaard' 'Bruno Labbadia' 'Norbert Meier' 'Sir Alex Ferguso
n'
```

In [149]:
```
information_cat("away_club_manager_name")
```

```
away_club_manager_name
Unique values: ['Kent Nielsen' 'Colin Todd' 'Troels Bech' 'Temur Shalamb
eridze'
 'Peter Maes' 'Auri Skarbalius' 'Ariel Jacobs' 'Jess Thorup'
 'Peter Rensen' 'Lars Sondergaard' 'Keld Bordinggaard' 'Valdas Urbonas'
 'Mirko Slomka' 'Kasper Hjulmand' 'Adrie Koster' 'Johnny Ma'
 'John van den Brom' 'Liam Buckley' 'Gertjan Verbeek' 'Enzo Scifo'
 'andre villas boas' 'Dieter Hecking' 'Peter Bosz' 'Jose Mota '
 'Sir Alex Ferguson' 'Markus Babbel' 'David Moyes' 'Steve Clarke'
 'Huub Stevens' 'Arsene Wenger' 'Jorge Jesus' 'Marco van Basten'
 'Rolando Maran' 'Pampos Christodoulou' 'Guus Hiddink' 'Orest Lenczyk'
 'Chris Hughton' 'Tony Pulis' 'Armin Veh' 'Jurgen Klopp'
 'Markus Weinzierl' 'Mario Been' 'Paul Lambert' 'Roberto Mancini'
 'Michael Laudrup' 'Thomas Schaaf' 'Ruud Brood' 'Glen Riddersholm'
 'Alan Pardew' 'Vladimir Petkovic' 'Jupp Heynckes' 'Mark Hughes'
 'Gert Heerkes' 'Nigel Adkins' 'Bruno Labbadia' 'Erwin Koeman'
 'Roberto Di Matteo' 'Manuel Pellegrini' 'JuanI Ignacio Martinez '
 'Stefano Colantuono' 'Mike BUskens' 'Zdena  Zeman' 'Angelo Alessio'
 'Massimiliano Allegri' nan "Martin O'Neill" 'Brendan Rodgers'
```

In [150]:
```
information_cat("away_club_position")
```

```
away_club_position
Unique values: [ 7. 12.  5. nan  9.  2. 11.  1.  6. 10. 16.  3.  8. 14.
4. 13. 15. 18.
 19. 17. 20. 21.]
Mode(most repeated values): 10.0
Missing values: 503
```

In [151]:
```
information_cat("home_club_position")
```

```
home_club_position
Unique values: [ 6.  1.  9. nan  3. 10. 12.  4.  7.  8.  2.  5. 15. 11. 2
0. 14. 16. 17.
 13. 19. 18.]
Mode(most repeated values): 3.0
Missing values: 503
```

In [152]:
```
# filling missing values
```

In [153]:
```
df_appear["away_club_position"].fillna(df_appear["away_club_position"].mode(
```

In [154]:
```
df_appear["away_club_position"].isnull().sum()
```

Out[154]: 0

In [155]:
```
df_appear["home_club_position"].fillna(df_appear["home_club_position"].mode(
```

In [156]:
```
df_appear["home_club_position"].isnull().sum()
```

Out[156]: 0

In [157]:
```
df_appear["away_club_manager_name"].fillna(df_appear["away_club_manager_name
```

In [158]:
```
df_appear["away_club_manager_name"].isnull().sum()
```

Out[158]: 0

In [159]:
```
df_appear["home_club_manager_name"].fillna(df_appear["home_club_manager_name
```

In [160]:
```
df_appear["home_club_manager_name"].isnull().sum()
```

Out[160]: 0

In [161]:
```
df_appear["agent_name"].fillna(df_appear["agent_name"].mode()[0], inplace =
```

In [162]:
```
df_appear["agent_name"].isnull().sum()
```

Out[162]: 0

In [163]:
```
df_appear["referee"].fillna(df_appear["referee"].mode()[0], inplace = True)
```

In [164]:
```
df_appear["referee"].isnull().sum()
```

Out[164]: 0

In [165]:
```
df_appear["home_club_name"].fillna(df_appear["home_club_name"].mode()[0], ir
```

In [166]:
```
df_appear["home_club_name"].isnull().sum()
```

Out[166]: 0

In [167]:
```
df_appear["away_club_name"].fillna(df_appear["away_club_name"].mode()[0], ir
```

In [168]:
```
df_appear["away_club_name"].isnull().sum()
```

Out[168]: 0

In [169]:
```
df_appear["foot"].fillna(df_appear["foot"].mode()[0], inplace = True)
```

In [170]:
```
df_appear["foot"].isnull().sum()
```

Out[170]: 0

In [171]:
```
df_appear["height_in_cm"].mean()
```

Out[171]: 184.26414557861813

In [172]:
```
df_appear["height_in_cm"].fillna(df_appear["height_in_cm"].mean(), inplace =
```

In [173]:
```
df_appear["height_in_cm"].isnull().sum()
```

Out[173]: 0

In [174]:
```
df_appear["market_value_in_eur"].fillna(df_appear["market_value_in_eur"].mea
```

In [175]:
```
df_appear["market_value_in_eur"].isnull().sum()
```

Out[175]: 0

In [176]:
```
df_appear["contract_expiration_date"].fillna(df_appear["contract_expiration_
```

In [177]:
```
df_appear["contract_expiration_date"].isnull().sum()
```

Out[177]: 0

In [178]:
```
df_appear["agent_name"].fillna(df_appear["agent_name"].mode()[0], inplace =
```

In [179]:
```
df_appear["agent_name"].isnull().sum()
```

Out[179]: 0

In [ ]:
```

```

In [180]:
```
df_appear["attendance"].mean()
```

Out[180]: 29820.04192139738

In [181]:
```
df_appear["attendance"].fillna(df_appear["attendance"].mean(), inplace = Tru
```

In [182]:
```
df_appear["attendance"].isnull().sum()
```

Out[182]: 0

In [183]:
```
df_appear.isnull().sum()
```

Out[183]:
```
appearance_id                  0
game_id                        0
player_id                      0
date_x                         0
player_name                    0
competition_id_x               0
yellow_cards                   0
red_cards                      0
goals                          0
assists                        0
minutes_played                 0
competition_id_y               0
season                         0
round                          0
date_y                         0
home_club_goals                0
away_club_goals                0
home_club_position             0
away_club_position             0
home_club_manager_name         0
away_club_manager_name         0
stadium                        0
attendance                     0
referee                        0
home_club_formation         3568
away_club_formation         3568
home_club_name                 0
away_club_name                 0
aggregate                      0
competition_type               0
name                           0
last_season                    0
current_club_id                0
player_code                    0
country_of_birth               0
date_of_birth                  0
sub_position                   0
position                       0
foot                           0
height_in_cm                   0
market_value_in_eur            0
highest_market_value_in_eur    0
contract_expiration_date       0
agent_name                     0
dtype: int64
```

In [ ]:
```

```

In [184]:
```
#removing duplicates
```

In [185]:
```
columns_to_exclude_for_df_appear = ["appearance_id", "game_id", "player_id",
df_appear = df_appear.drop_duplicates(subset = [col for col in df_appear.col
```

In [186]:
```
df_appear
```

Out[186]:

	appearance_id	game_id	player_id	date_x	player_name	competition_id_x	yellow_
0	2224728_119169	2224728	119169	2012-07-13	Aron Johannsson	DK1	
1	2224732_161244	2224732	161244	2012-07-14	Conor O'Brien	DK1	
2	2224729_39467	2224729	39467	2012-07-15	Clarence Goodson	DK1	
3	2232104_119169	2232104	119169	2012-07-19	Aron Johannsson	ELQ	
4	2219794_39475	2219794	39475	2012-07-22	Sacha Kljestan	BESC	
...	...	...	...	...	...	...	
3563	3415291_537467	3415291	537467	2020-09-26	Joseph Efford	BE1	
3564	3415296_367423	3415296	367423	2020-	Chris Durkin	BE1	

In [ ]:
```
#Coverting the data into csv file
```

In [200]:
```
df_appear.to_csv("df_appear_cleaned.csv")
```

In [ ]:
```

```

for df_lineup

In [187]:
```
df_lineup
```

Out[187]:

	game_lineups_id	game_id	type	number	player_id	player_n
0	f2570d1504fc02f4b6c7608e8dcf89a3	4087925	substitutes	34	242284	E Hol
1	f5f0da93ea8e1d8bdd799658e7c8f7cb	4087928	starting_lineup	13	145466	Tim R
2	31a4d12ec23d604779d909d26c1b5410	4087929	substitutes	26	578539	(Rich
3	776dcbef98651450db76723cb7e3b4df	4087935	substitutes	26	578539	(Rich
4	6a35ef7495303f29e7f85dbd54547fb1	4087936	starting_lineup	13	145466	Tim R
...	...	...	...	...	...	
214	ec3d266094f99ca0a8847de827e37105	4194152	starting_lineup	7	504215	Giov R
215	667840cda9bdf3b0344b8e99b306cf38	4194152	starting_lineup	23	124732	John Ant Br
216	1c5d2f60ee777760f8a757aa10c42bb1	4194154	starting_lineup	13	103064	Terr I
217	99032084fd00ffbfee52c541a9f960ab	4204000	substitutes	14	315762	Luca .
218	5c9eaf6ebb621d43a0d6fd6a9e607ef9	4220942	substitutes	13	145466	Tim R

219 rows × 41 columns

In [188]:
```
def cat_num_data_types(df_lineup):
    categorical = []
    continuous = []
    for column in df_lineup.columns:
        if df_lineup[column].nunique() < 30:
            categorical.append(column)
        else:
            continuous.append(column)

    return categorical, continuous

categorical, continuous = cat_num_data_types(df_lineup)

#Tabulate is a package used to print the list, dict or any data sets in a pr

from tabulate import tabulate
table = [categorical, continuous]
print(tabulate({"Categorical":categorical, "Continuous":continuous}, headers
```

```
categorical                 continuous
--------------------------  ---------------
type                        game_lineups_id
number                      game_id
player_id
player_name
team_captain
position_x
competition_id
season
round
date
home_club_goals
away_club_goals
home_club_position
away_club_position
home_club_manager_name
away_club_manager_name
stadium
attendance
referee
home_club_formation
away_club_formation
home_club_name
away_club_name
aggregate
competition_type
name
last_season
current_club_id
player_code
country_of_birth
date_of_birth
sub_position
position_y
foot
height_in_cm
market_value_in_eur
highest_market_value_in_eur
contract_expiration_date
agent_name
```

In [189]:
```
df_lineup.isnull().sum()
```

Out[189]:
```
game_lineups_id                 0
game_id                         0
type                            0
number                          0
player_id                       0
player_name                     0
team_captain                    0
position_x                      0
competition_id                219
season                        219
round                         219
date                          219
home_club_goals               219
away_club_goals               219
home_club_position            219
away_club_position            219
home_club_manager_name        219
away_club_manager_name        219
stadium                       219
attendance                    219
referee                       219
home_club_formation           219
away_club_formation           219
home_club_name                219
away_club_name                219
aggregate                     219
competition_type              219
name                            0
last_season                     0
current_club_id                 0
player_code                     0
country_of_birth                0
date_of_birth                   0
sub_position                    0
position_y                      0
foot                            0
height_in_cm                    0
market_value_in_eur             1
highest_market_value_in_eur     0
contract_expiration_date        9
agent_name                     67
dtype: int64
```

In [192]:
```
#Separating the data into categorical and continuous
```

In [193]:
```
def for_cat(col):           #categorical
    print(col)
    print("Unique values:", df_lineup[col].unique())
    print("Mode(most repeated values):", df_lineup[col].mode()[0])
    print("Missing values:",df_lineup[col].isnull().sum())
```

In [194]:
```
for_cat("agent_name")
```

```
agent_name
Unique values: ['SBM' 'CAA Stellar' nan 'acta7' 'TrueSports GmbH' 'ROGON'
'Gestifute'
 'in4 sports' 'Avid Sports Group' 'Wasserman' 'Joes Blakborn'
 'BS Group - BS Law' 'Field Management' 'TOP Agency' 'PROSPORT Management'
 'PRO FC' 'Unique Sports Group' 'YMU Management Ltd.']
Mode(most repeated values): Wasserman
Missing values: 67
```

In [195]:
```
df_lineup["agent_name"].fillna(df_lineup["agent_name"].mode()[0], inplace =
```

In [196]:
```
df_lineup["agent_name"].isnull().sum()
```

Out[196]: 0

In [197]:
```
df_lineup["market_value_in_eur"].mean()
```

Out[197]: 7827522.9357798165

In [198]:
```
df_lineup["market_value_in_eur"].fillna(df_lineup["market_value_in_eur"].mea
```

In [199]:
```
df_lineup["market_value_in_eur"].isnull().sum()
```

Out[199]: 0

In [ ]:
```

```

In [200]:
```
df_lineup["contract_expiration_date"].fillna(df_lineup["contract_expiration_
```

In [201]:
```
df_lineup["contract_expiration_date"].isnull().sum()
```

Out[201]: 0

In [202]:
```
df_lineup.isnull().sum()
```

Out[202]:
```
game_lineups_id                0
game_id                        0
type                           0
number                         0
player_id                      0
player_name                    0
team_captain                   0
position_x                     0
competition_id               219
season                       219
round                        219
date                         219
home_club_goals              219
away_club_goals              219
home_club_position           219
away_club_position           219
home_club_manager_name       219
away_club_manager_name       219
stadium                      219
attendance                   219
referee                      219
home_club_formation          219
away_club_formation          219
home_club_name               219
away_club_name               219
aggregate                    219
competition_type             219
name                           0
last_season                    0
current_club_id                0
player_code                    0
country_of_birth               0
date_of_birth                  0
sub_position                   0
position_y                     0
foot                           0
height_in_cm                   0
market_value_in_eur            0
highest_market_value_in_eur    0
contract_expiration_date       0
agent_name                     0
dtype: int64
```

In [ ]:
```

```

In [203]:
```
df_lineup
```

...

In [204]:
```
#Converting data into csv file
```

In [32]:
```
df_lineup.to_csv("df_lineup_cleaned.csv")
```

In [ ]: