In [1]:
```
#import library
import pandas as pd
import numpy  as np
```

In [2]:
```
#read data
dt = pd.read_csv("top250.csv")
dt_fix = pd.read_csv('top250.csv')
print(dt.head(10))
```

```
              Name          Position  Age       Team_from     League_from  \
0         Luís Figo     Right Winger   27    FC Barcelona          LaLiga
1     Hernán Crespo    Centre-Forward   25           Parma        Serie A
2     Marc Overmars       Left Winger   27         Arsenal  Premier League
3  Gabriel Batistuta    Centre-Forward   31      Fiorentina        Serie A
4    Nicolas Anelka    Centre-Forward   21     Real Madrid          LaLiga
5     Rio Ferdinand       Centre-Back   22        West Ham  Premier League
6   Flávio Conceicao  Central Midfield   26  Dep. La Coruña          LaLiga
7    Savo Milosevic    Centre-Forward   26   Real Zaragoza          LaLiga
8   David Trézéguet    Centre-Forward   22          Monaco        Ligue 1
9     Claudio López    Centre-Forward   25     Valencia CF          LaLiga

       Team_to       League_to     Season  Market_value  Transfer_fee
0   Real Madrid          LaLiga  2000-2001           NaN      60000000
1         Lazio        Serie A  2000-2001           NaN      56810000
2  FC Barcelona          LaLiga  2000-2001           NaN      40000000
3      AS Roma        Serie A  2000-2001           NaN      36150000
4     Paris SG        Ligue 1  2000-2001           NaN      34500000
5         Leeds  Premier League  2000-2001           NaN      26000000
6   Real Madrid          LaLiga  2000-2001           NaN      25000000
7         Parma        Serie A  2000-2001           NaN      25000000
8      Juventus        Serie A  2000-2001           NaN      23240000
9         Lazio        Serie A  2000-2001           NaN      23000000
```

In [3]:
```
#review data
print(dt.shape)
dt.sort_values(by = ['Season','Transfer_fee'],ascending = [True,False])
dt_fix.sort_values(by=['Season','Transfer_fee'],ascending = [True,False])
```

```
(4700, 10)
```

Out[3]:

	Name	Position	Age	Team_from	League_from	Team_to	League_to	Season	Market_value	Transfer_fee
0	Luís Figo	Right Winger	27	FC Barcelona	LaLiga	Real Madrid	LaLiga	2000-2001	NaN	60000000
1	Hernán Crespo	Centre-Forward	25	Parma	Serie A	Lazio	Serie A	2000-2001	NaN	56810000
2	Marc Overmars	Left Winger	27	Arsenal	Premier League	FC Barcelona	LaLiga	2000-2001	NaN	40000000
3	Gabriel Batistuta	Centre-Forward	31	Fiorentina	Serie A	AS Roma	Serie A	2000-2001	NaN	36150000
4	Nicolas Anelka	Centre-Forward	21	Real Madrid	LaLiga	Paris SG	Ligue 1	2000-2001	NaN	34500000
...	...	...	...	...	...	...	...	...	...	...
4695	Jasmin Kurtic	Attacking Midfield	29	Atalanta	Serie A	SPAL	Serie A	2018-2019	5000000.0	4800000
4696	Tchê Tchê	Central Midfield	25	Palmeiras	Série A	Dynamo Kyiv	Premier Liga	2018-2019	3000000.0	4800000
4697	Silvan Widmer	Right-Back	25	Udinese Calcio	Serie A	FC Basel	Super League	2018-2019	8500000.0	4500000
4698	Yuya Osako	Second Striker	28	1. FC Köln	2.Bundesliga	Werder Bremen	1.Bundesliga	2018-2019	4500000.0	4500000
4699	Kyle Bartley	Centre-Back	27	Swansea	Championship	West Brom	Championship	2018-2019	3500000.0	4500000

4700 rows × 10 columns

In [4]:
```
#check for duplication
dup = dt.duplicated()
duprow = dt[dup]
print(duprow)
```

```
Empty DataFrame
Columns: [Name, Position, Age, Team_from, League_from, Team_to, League_to, Season, Market_value, Transfer_fee]
Index: []
```

In [5]:
```
#check null
print(dt.isnull().sum())
```

```
Name              0
Position          0
Age               0
Team_from         0
League_from       0
Team_to           0
League_to         0
Season            0
Market_value    1260
Transfer_fee       0
dtype: int64
```

In [6]:
```
#check each column
```

In [47]:
```
#check 'Name' column and replace abbreviation
##Alex
print(dt['Name'].describe())
dt_fix.loc[[321,357,1092],'Name']   = "Alexsandro de Souza"
dt_fix.loc[1007,'Name']             = "Alex de Oliveira Nascimento"
dt_fix.loc[[2139,2584,3099],'Name'] = "Alexandre Raphael Meschini"
dt_fix.loc[2880,'Name']             = "Alex Rodrigo Dias da Costa"
dt_fix.loc[969,'Name']              = "Fernando Santos"
dt_fix.loc[2859,'Name']             = "Fernando Menegazzo"
dt_fix.loc[[3292,3876,4074],'Name'] = "Fernando Lucas Martins"
dt_fix.loc[3509,'Name']             = "Fernando Francisco Reges"
dt_fix.loc[4662,'Name']             = "Fernando dos Santos Pedro"
dt_fix.loc[[284,510,746,2644],'Name'] = "Adriano Leite Ribeiro"
dt_fix.loc[[3123,3247,3784,4233],'Name'] = "José Paulo Bezerra Maciel Júnior"
dt_fix = dt_fix.replace("Émerson","Émerson Ferreira da Rosa")
dt_fix = dt_fix.replace("Sokratis", "Sokratis Papastathopoulos")
print("\n final result:")
print(dt_fix['Name'].describe())
print(dt_fix['Name'].value_counts())
```

```
count     4700
unique    3104
top       Alex
freq         8
Name: Name, dtype: object

 final result:
count             4700
unique            3113
top       Peter Crouch
freq                 7
Name: Name, dtype: object
Peter Crouch                7
Alberto Gilardino           6
Robbie Keane                6
Sokratis Papastathopoulos   6
Émerson Ferreira da Rosa    6
                           ..
Fábio Simplício             1
Celsinho                    1
Ricardo Rocha               1
Benoît Assou-Ekotto         1
Kyle Bartley                1
Name: Name, Length: 3113, dtype: int64
```

In [27]:
```
print(dt['Position'].describe())
print(dt['Position'].unique())
```

```
count               4700
unique                17
top       Centre-Forward
freq                1218
Name: Position, dtype: object
['Right Winger' 'Centre-Forward' 'Left Winger' 'Centre-Back'
 'Central Midfield' 'Attacking Midfield' 'Defensive Midfield'
 'Second Striker' 'Goalkeeper' 'Right-Back' 'Left Midfield' 'Left-Back'
 'Right Midfield' 'Forward' 'Sweeper' 'Defender' 'Midfielder']
```

In [28]:
```
print(dt['Age'].describe())
#clean age = 0 by using mean
mean_age = int(dt['Age'].mean())
dt_fix['Age'] = dt_fix['Age'].fillna(mean_age)
dt_fix.loc[dt_fix['Age'] == 0,"Age"] = mean_age
print("\nfinal_data:")
print(dt_fix['Age'].describe())
```

```
count    4700.000000
mean       24.338723
std         3.230809
min         0.000000
25%        22.000000
50%        24.000000
75%        27.000000
max        35.000000
Name: Age, dtype: float64

final_data:
count    4700.00000
mean       24.34383
std         3.21124
min        15.00000
25%        22.00000
50%        24.00000
75%        27.00000
max        35.00000
Name: Age, dtype: float64
```

In [49]:
```
print(dt[['Team_from','Team_to']].describe())
```

```
       Team_from Team_to
count       4700    4700
unique       570     325
top        Inter   Inter
freq          68      97
```

In [30]:
```
print(dt[['League_from','League_to']].describe())
```

```
           League_from       League_to
count             4700            4700
unique             118              65
top     Premier League  Premier League
freq               608            1256
```

In [31]:
```
print(dt['Season'].describe())
```

```
count          4700
unique           19
top       2001-2002
freq            250
Name: Season, dtype: object
```

In [32]:
```
print(dt['Market_value'].describe())
```

```
count    3.440000e+03
mean     8.622469e+06
std      8.795181e+06
min      5.000000e+04
25%      3.500000e+06
50%      6.000000e+06
75%      1.000000e+07
max      1.200000e+08
Name: Market_value, dtype: float64
```

In [33]:
```
print(dt['Transfer_fee'].describe())
```

```
count    4.700000e+03
mean     9.447586e+06
std      1.043772e+07
min      8.250000e+05
25%      4.000000e+06
50%      6.500000e+06
75%      1.082000e+07
max      2.220000e+08
Name: Transfer_fee, dtype: float64
```

In [34]:
```
#eliminate irrelevant information
if 'Market_value' in dt_fix.keys():
    #market_value is old and irrelevant
    dt_fix = dt_fix.drop(columns = 'Market_value')
```

In [35]:
```
print(dt_fix.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4700 entries, 0 to 4699
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   Name           4700 non-null   object
 1   Position       4700 non-null   object
 2   Age            4700 non-null   int64 
 3   Team_from      4700 non-null   object
 4   League_from    4700 non-null   object
 5   Team_to        4700 non-null   object
 6   League_to      4700 non-null   object
 7   Season         4700 non-null   object
 8   Transfer_fee   4700 non-null   int64 
dtypes: int64(2), object(7)
memory usage: 330.6+ KB
None
```

In [36]:
```
#review final result
dt_fix.head(10)
```

Out[36]:

	Name	Position	Age	Team_from	League_from	Team_to	League_to	Season	Transfer_fee
0	Luís Figo	Right Winger	27	FC Barcelona	LaLiga	Real Madrid	LaLiga	2000-2001	60000000
1	Hernán Crespo	Centre-Forward	25	Parma	Serie A	Lazio	Serie A	2000-2001	56810000
2	Marc Overmars	Left Winger	27	Arsenal	Premier League	FC Barcelona	LaLiga	2000-2001	40000000
3	Gabriel Batistuta	Centre-Forward	31	Fiorentina	Serie A	AS Roma	Serie A	2000-2001	36150000
4	Nicolas Anelka	Centre-Forward	21	Real Madrid	LaLiga	Paris SG	Ligue 1	2000-2001	34500000
5	Rio Ferdinand	Centre-Back	22	West Ham	Premier League	Leeds	Premier League	2000-2001	26000000
6	Flávio Conceicao	Central Midfield	26	Dep. La Coruña	LaLiga	Real Madrid	LaLiga	2000-2001	25000000
7	Savo Milosevic	Centre-Forward	26	Real Zaragoza	LaLiga	Parma	Serie A	2000-2001	25000000
8	David Trézéguet	Centre-Forward	22	Monaco	Ligue 1	Juventus	Serie A	2000-2001	23240000
9	Claudio López	Centre-Forward	25	Valencia CF	LaLiga	Lazio	Serie A	2000-2001	23000000

In [37]:
```
dt_fix.to_csv('top250c.csv', index = False)
```

In [ ]:
```

```