In [104]:

```
#import the libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [31]:

```
#Load dataset
epl_df=pd.read_csv('C:\\Users\\roshn\\OneDrive\\Desktop\\Data Science\\EPL_20_21.csv')
epl_df.head()
```

Out[31]:

	Name	Club	Nationality	Position	Age	Matches	Starts	Mins	Goals	Assists	Passes
0	Mason Mount	Chelsea	ENG	MF,FW	21	36	32	2890	6	5	1881
1	Edouard Mendy	Chelsea	SEN	GK	28	31	31	2745	0	0	1007
2	Timo Werner	Chelsea	GER	FW	24	35	29	2602	6	8	826
3	Ben Chilwell	Chelsea	ENG	DF	23	27	27	2286	3	5	1806
4	Reece James	Chelsea	ENG	DF	20	32	25	2373	1	2	1987

In [32]:

```
epl_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 532 entries, 0 to 531
Data columns (total 18 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   Name                   532 non-null    object 
 1   Club                   532 non-null    object 
 2   Nationality            532 non-null    object 
 3   Position               532 non-null    object 
 4   Age                    532 non-null    int64  
 5   Matches                532 non-null    int64  
 6   Starts                 532 non-null    int64  
 7   Mins                   532 non-null    int64  
 8   Goals                  532 non-null    int64  
 9   Assists                532 non-null    int64  
 10  Passes_Attempted       532 non-null    int64  
 11  Perc_Passes_Completed  532 non-null    float64
 12  Penalty_Goals          532 non-null    int64  
 13  Penalty_Attempted      532 non-null    int64  
 14  xG                     532 non-null    float64
 15  xA                     532 non-null    float64
 16  Yellow_Cards           532 non-null    int64  
 17  Red_Cards              532 non-null    int64  
dtypes: float64(3), int64(11), object(4)
memory usage: 74.9+ KB
```

In [33]:

```
epl_df.describe() #only works on num values
```

Out[33]:

	Age	Matches	Starts	Mins	Goals	Assists	Passes_Atte
count	532.000000	532.000000	532.000000	532.000000	532.000000	532.000000	532.000000
mean	25.500000	19.535714	15.714286	1411.443609	1.853383	1.287594	717.750000
std	4.319404	11.840459	11.921161	1043.171856	3.338009	2.095191	631.372522
min	16.000000	1.000000	0.000000	1.000000	0.000000	0.000000	0.000000
25%	22.000000	9.000000	4.000000	426.000000	0.000000	0.000000	171.500000
50%	26.000000	21.000000	15.000000	1345.000000	1.000000	0.000000	573.500000
75%	29.000000	30.000000	27.000000	2303.500000	2.000000	2.000000	1129.500000
max	38.000000	38.000000	38.000000	3420.000000	23.000000	14.000000	3214.000000

In [34]:

```
epl_df.isna().sum() #checking null values
```

Out[34]:

```
Name                     0
Club                     0
Nationality              0
Position                 0
Age                      0
Matches                  0
Starts                   0
Mins                     0
Goals                    0
Assists                  0
Passes_Attempted         0
Perc_Passes_Completed    0
Penalty_Goals            0
Penalty_Attempted        0
xG                       0
xA                       0
Yellow_Cards             0
Red_Cards                0
dtype: int64
```

In [37]:

```
#create 2 new columns
epl_df['MinsPerMatch']=(epl_df['Mins']/epl_df['Matches']).astype(int)
epl_df['GoalsPerMatch']=(epl_df['Goals']/epl_df['Matches']).astype(float)
print(epl_df.head())
```

```
            Name     Club Nationality Position  Age  Matches  Starts  Mins  \
0    Mason Mount  Chelsea         ENG    MF,FW   21       36      32  2890
1  Edouard Mendy  Chelsea         SEN       GK   28       31      31  2745
2    Timo Werner  Chelsea         GER       FW   24       35      29  2602
3   Ben Chilwell  Chelsea         ENG       DF   23       27      27  2286
4    Reece James  Chelsea         ENG       DF   20       32      25  2373

   Goals  Assists  Passes_Attempted  Perc_Passes_Completed  Penalty_Goals  \
0      6        5              1881                   82.3              1
1      0        0              1007                   84.6              0
2      6        8               826                   77.2              0
3      3        5              1806                   78.6              0
4      1        2              1987                   85.0              0

   Penalty_Attempted    xG    xA  Yellow_Cards  Red_Cards  MinsPerMatch  \
0                  1  0.21  0.24             2          0            80
1                  0  0.00  0.00             2          0            88
2                  0  0.41  0.21             2          0            74
3                  0  0.10  0.11             3          0            84
4                  0  0.06  0.12             3          0            74
```

```
   GoalsPerMatch
0       0.166667
1       0.000000
2       0.171429
3       0.111111
4       0.031250
```

In [38]:

```
epl_df.head()
```

Out[38]:

	Name	Club	Nationality	Position	Age	Matches	Starts	Mins	Goals	Assists	Passes
0	Mason Mount	Chelsea	ENG	MF,FW	21	36	32	2890	6	5	1881
1	Edouard Mendy	Chelsea	SEN	GK	28	31	31	2745	0	0	1007
2	Timo Werner	Chelsea	GER	FW	24	35	29	2602	6	8	826
3	Ben Chilwell	Chelsea	ENG	DF	23	27	27	2286	3	5	1806
4	Reece James	Chelsea	ENG	DF	20	32	25	2373	1	2	1987

In [41]:

```
#Total Goals
Total_Goals=epl_df['Goals'].sum()
print(Total_Goals)
```

```
986
```

In [42]:

```
#Penalty goals
Total_Penalty_Goals=epl_df['Penalty_Goals'].sum()
print(Total_Penalty_Goals)
```

```
102
```

In [43]:

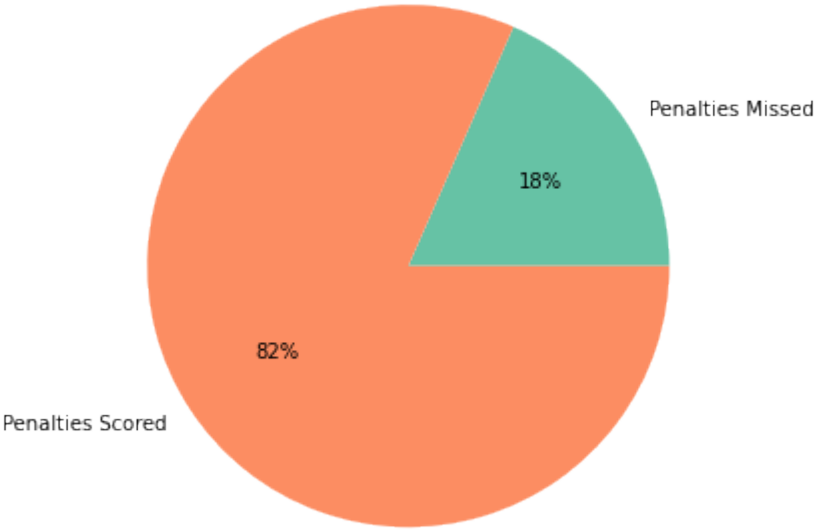
```
#Penalty Attempts
Total_PenaltyAttempts=epl_df['Penalty_Attempted'].sum()
print(Total_PenaltyAttempts)
```

```
125
```

In [58]:

```
#Pie chart for penalties missed vs scored

plt.figure(figsize=(13,6))
pl_not_scored=epl_df['Penalty_Attempted'].sum() - Total_Penalty_Goals
data=[pl_not_scored,Total_Penalty_Goals]
labels=['Penalties Missed','Penalties Scored']
color=sns.color_palette('Set2')
plt.pie(data,labels=labels,colors=color,autopct='%.0f%%')
plt.show()
```



In [59]:

```
#unique positions
epl_df['Position'].unique()
```

Out[59]:

```
array(['MF,FW', 'GK', 'FW', 'DF', 'MF', 'FW,MF', 'FW,DF', 'DF,MF',
       'MF,DF', 'DF,FW'], dtype=object)
```

In [61]:

```
#total number of forward players
epl_df[epl_df['Position']=='FW']
```

Out[61]:

	Name	Club	Nationality	Position	Age	Matches	Starts	Mins	Goals	Assists	P
2	Timo Werner	Chelsea	GER	FW	24	35	29	2602	6	8	8
16	Tammy Abraham	Chelsea	ENG	FW	22	22	12	1040	6	1	2
19	Olivier Giroud	Chelsea	FRA	FW	33	17	8	748	4	0	2
23	Ruben Loftus-Cheek	Chelsea	ENG	FW	24	1	1	60	0	0	1
30	Raheem Sterling	Manchester City	ENG	FW	25	31	28	2536	10	7	1
...	...	...	...	...	...	...	...	...	...	...	.
516	Oliver Burke	Sheffield United	SCO	FW	23	25	14	1269	1	1	2
518	Oliver McBurnie	Sheffield United	SCO	FW	24	23	12	1324	1	0	4
519	Rhian Brewster	Sheffield United	ENG	FW	20	27	12	1128	0	0	2
523	Billy Sharp	Sheffield United	ENG	FW	34	16	7	735	3	0	1
526	Daniel Jebbison	Sheffield United	ENG	FW	17	4	3	284	1	0	3

81 rows × 20 columns

In [65]:

```
#PLAYERS from different nations
np.size((epl_df['Nationality'].unique()))
```

Out[65]:

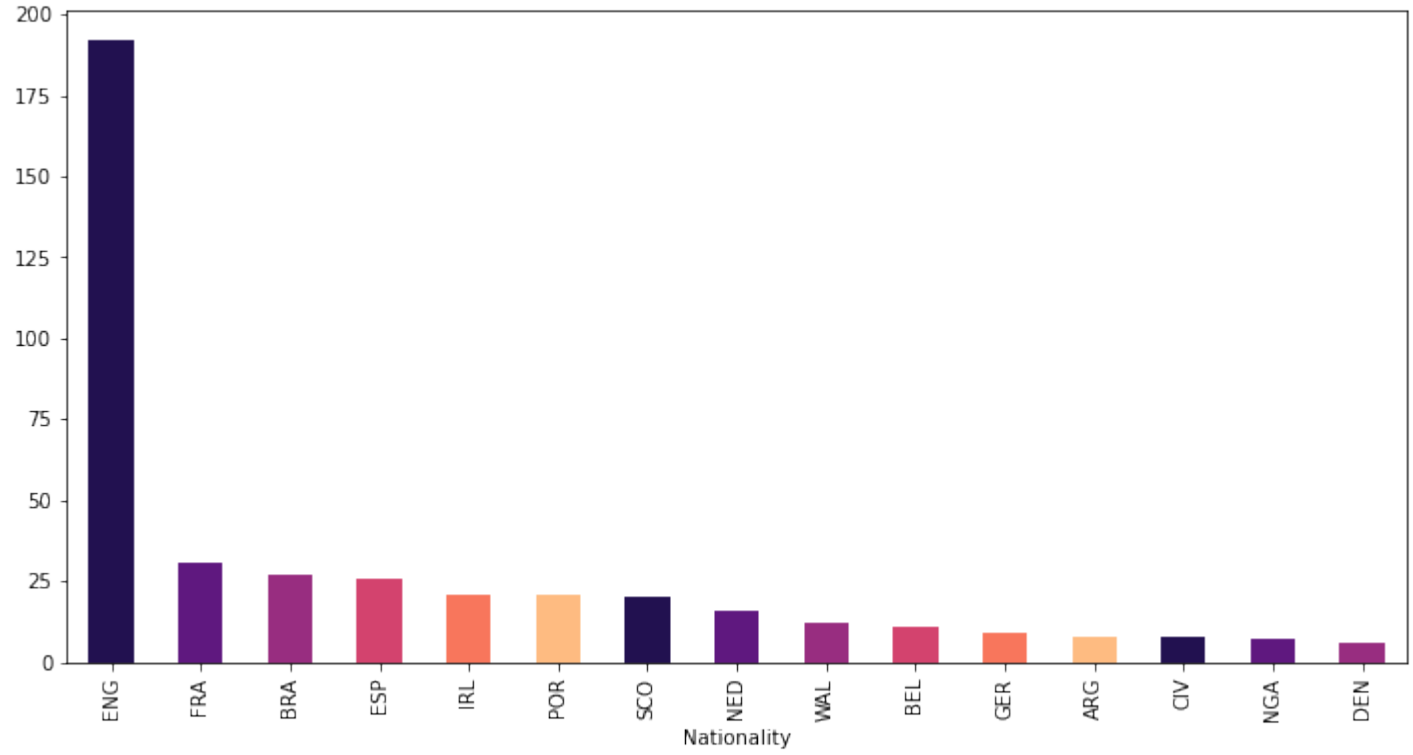
```
59
```

In [70]:

```
#most of players from which country?
nationality=epl_df.groupby('Nationality').size().sort_values(ascending= False)
nationality.head(15).plot(kind= 'bar',figsize=(12,6),color= sns.color_palette('magma'))
```

Out[70]:

```
<AxesSubplot:xlabel='Nationality'>
```

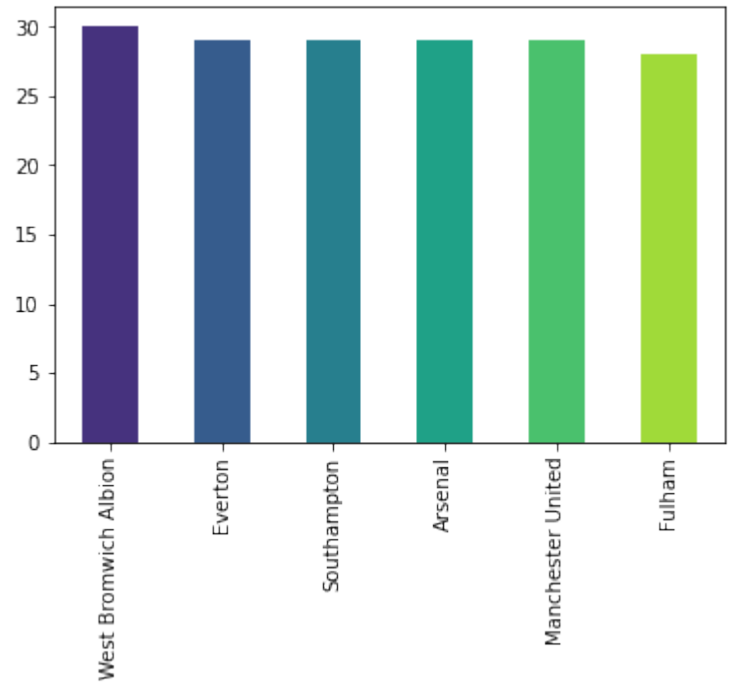


In [75]:

```
#clubs with maximum players in their squad
epl_df['Club'].value_counts().nlargest(6).plot(kind='bar',color= sns.color_palette('viridis'))
```

Out[75]:

```
<AxesSubplot:>
```

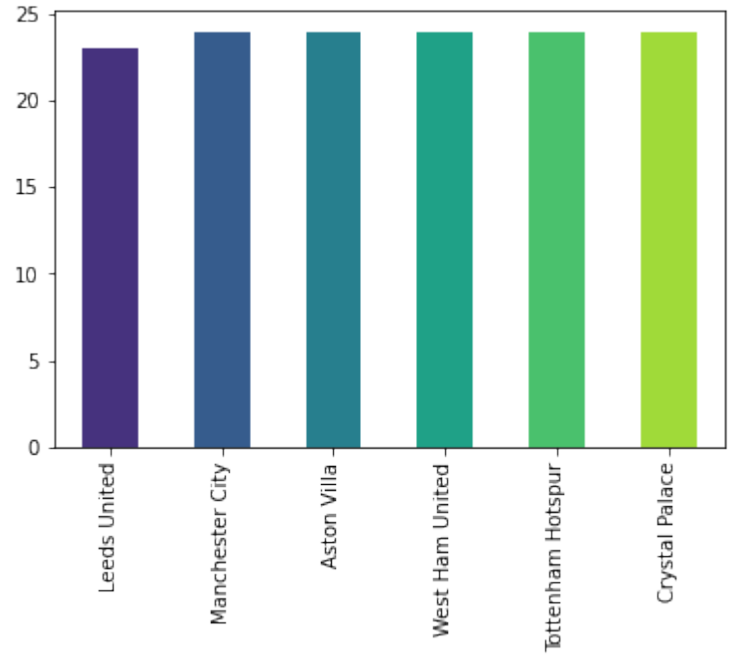


In [76]:

```
#clubs with least players in their squad
epl_df['Club'].value_counts().nsmallest(6).plot(kind='bar',color= sns.color_palette('viridis')
```

Out[76]:

```
<AxesSubplot:>
```

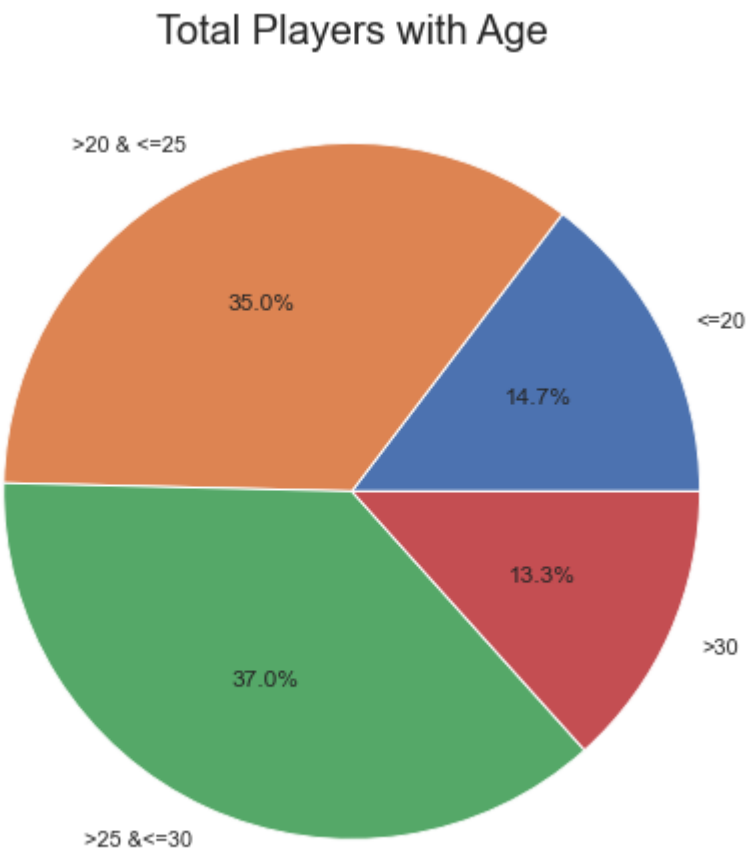


In [79]:

```
#players based on the age group
under20=epl_df[epl_df['Age']<=20]
age20_25=epl_df[(epl_df['Age']>20) & (epl_df['Age']<=25)]
age25_30=epl_df[(epl_df['Age']>25) & (epl_df['Age']<=30)]
above30=epl_df[epl_df['Age']>30]
```

In [134]:

```
x=np.array([under20['Name'].count(),age20_25['Name'].count(),age25_30['Name'].count(),above30[
mylabels=['<=20','>20 & <=25','>25 &<=30','>30']
plt.title('Total Players with Age',fontsize=20)
plt.pie(x,labels=mylabels,autopct="%.1f%%")
plt.show()
```

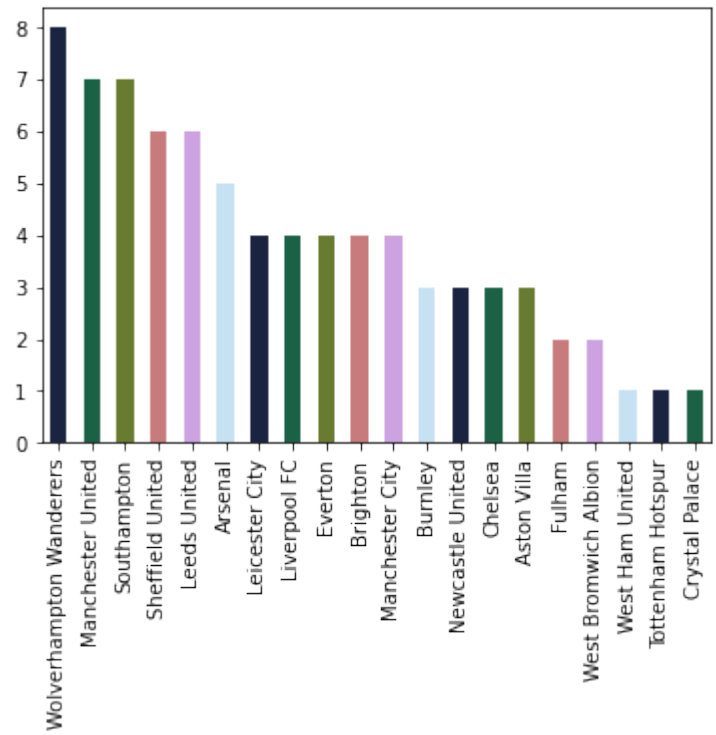


In [87]:

```
#total under 20 players in each club
Players_under_20=epl_df[epl_df['Age']<=20]
Players_under_20['Club'].value_counts().plot(kind='bar',color=sns.color_palette('cubehelix'))
```

Out[87]:

```
<AxesSubplot:>
```



In [88]:

```
#under 20 players in ManchesterUnited
Players_under_20[Players_under_20['Club']=='Manchester United']
```

Out[88]:

	Name	Club	Nationality	Position	Age	Matches	Starts	Mins	Goals	Assists	P
61	Mason Greenwood	Manchester United	ENG	FW	18	31	21	1822	7	2	7
72	Brandon Williams	Manchester United	ENG	DF	19	4	2	188	0	0	1
73	Amad Diallo	Manchester United	CIV	FW	18	3	2	166	0	1	6
74	Anthony Elanga	Manchester United	SWE	FW	18	2	2	155	1	0	5
76	Shola Shoretire	Manchester United	ENG	FW	16	2	0	11	0	0	8
78	Hannibal Mejbri	Manchester United	FRA	MF	17	1	0	9	0	0	3
79	William Thomas Fish	Manchester United	ENG	DF	17	1	0	1	0	0	1

In [89]:

```
#under 20 players in chelsea
Players_under_20[Players_under_20['Club']=='Chelsea']
```

Out[89]:

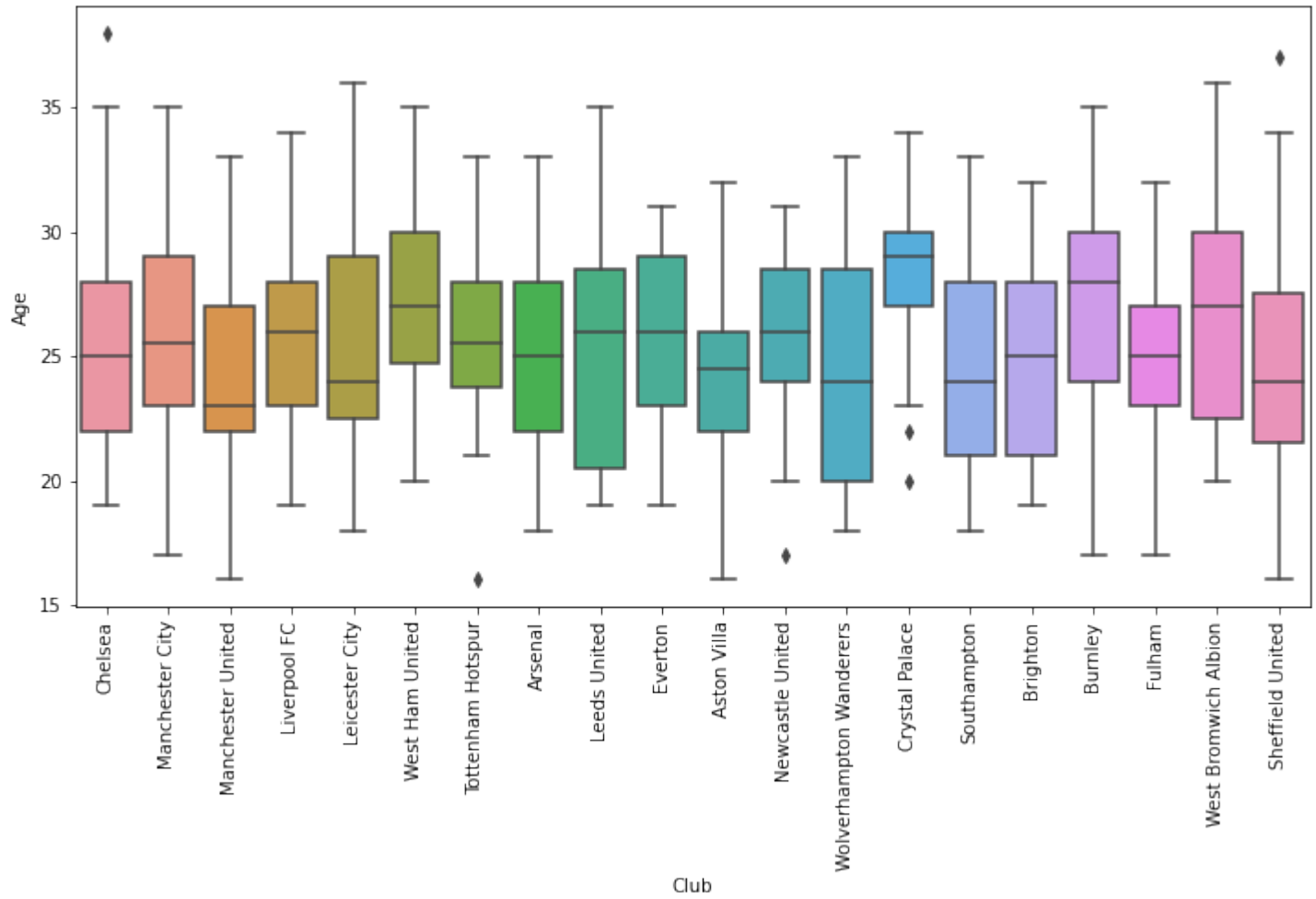
	Name	Club	Nationality	Position	Age	Matches	Starts	Mins	Goals	Assists	Passes
4	Reece James	Chelsea	ENG	DF	20	32	25	2373	1	2	1987
18	Callum Hudson-Odoi	Chelsea	ENG	FW,DF	19	23	10	1059	2	3	659
21	Billy Gilmour	Chelsea	SCO	MF	19	5	3	261	0	0	215

In [90]:

```
#Average age of players in each club
plt.figure(figsize=(12,6))
sns.boxplot(x='Club',y='Age',data=epl_df)
plt.xticks(rotation=90)
```

Out[90]:

```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16,
        17, 18, 19]),
 [Text(0, 0, 'Chelsea'),
  Text(1, 0, 'Manchester City'),
  Text(2, 0, 'Manchester United'),
  Text(3, 0, 'Liverpool FC'),
  Text(4, 0, 'Leicester City'),
  Text(5, 0, 'West Ham United'),
  Text(6, 0, 'Tottenham Hotspur'),
  Text(7, 0, 'Arsenal'),
  Text(8, 0, 'Leeds United'),
  Text(9, 0, 'Everton'),
  Text(10, 0, 'Aston Villa'),
  Text(11, 0, 'Newcastle United'),
  Text(12, 0, 'Wolverhampton Wanderers'),
  Text(13, 0, 'Crystal Palace'),
  Text(14, 0, 'Southampton'),
  Text(15, 0, 'Brighton'),
  Text(16, 0, 'Burnley'),
  Text(17, 0, 'Fulham'),
  Text(18, 0, 'West Bromwich Albion'),
  Text(19, 0, 'Sheffield United')])
```



In [92]:

```
num_player=epl_df.groupby('Club').size()
data=(epl_df.groupby('Club')['Age'].sum())/num_player
data.sort_values(ascending=False)
```

Out[92]:

```
Club
Crystal Palace              28.333333
West Ham United             27.500000
Burnley                     27.040000
West Bromwich Albion        26.766667
Newcastle United            26.074074
Manchester City             25.708333
Tottenham Hotspur           25.625000
Chelsea                     25.592593
Leicester City              25.592593
Liverpool FC                25.571429
Everton                     25.413793
Leeds United                25.347826
Fulham                      25.035714
Arsenal                     24.965517
Sheffield United            24.814815
Brighton                    24.555556
Wolverhampton Wanderers     24.444444
Aston Villa                 24.291667
Southampton                 24.137931
Manchester United           23.862069
dtype: float64
```

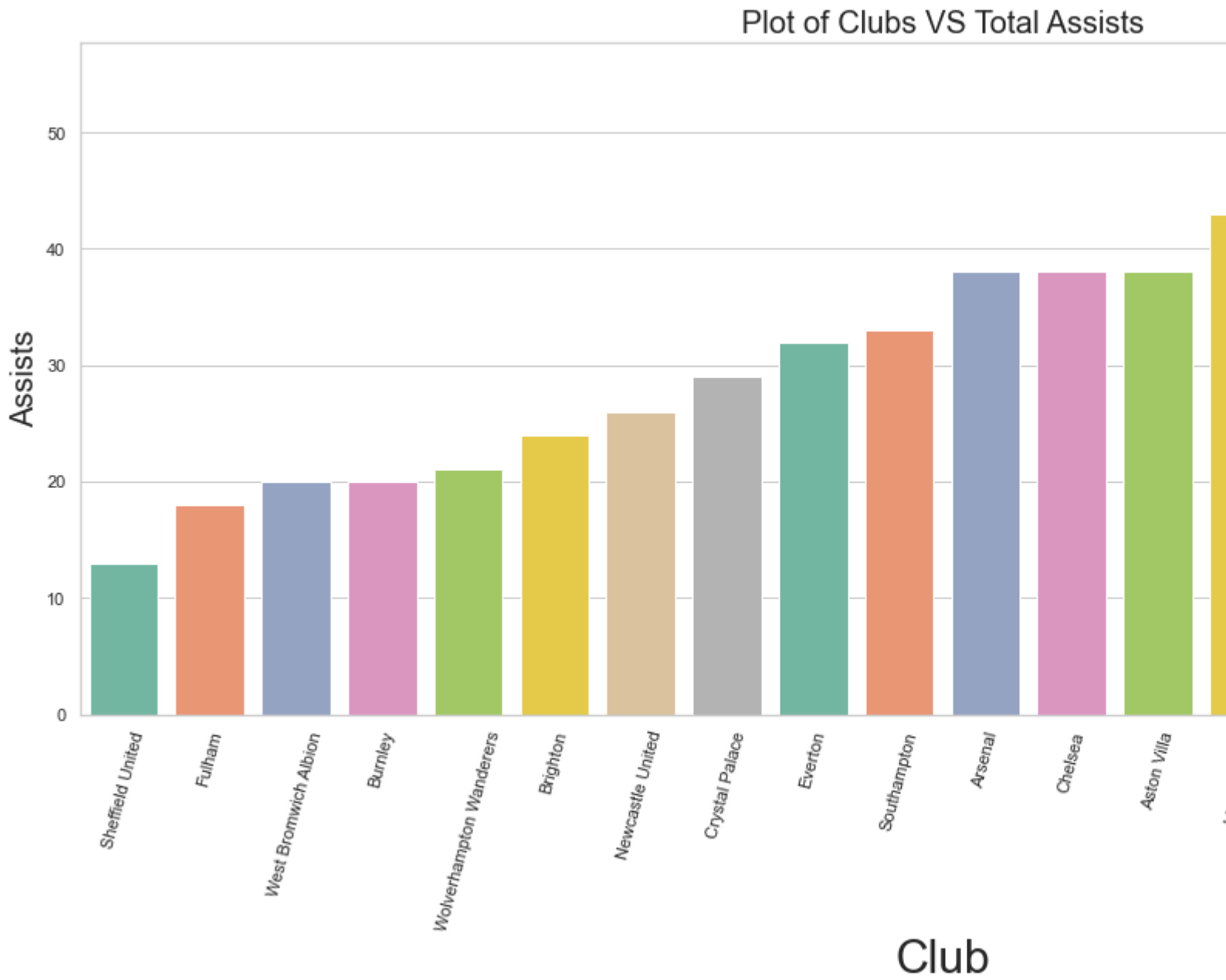
In [115]:

```
#Total assists from each club
import pandas as pd
Assists_by_clubs = pd.DataFrame(epl_df.groupby('Club',as_index=False)['Assists'].sum())
```

```
sns.set_theme(style="whitegrid",color_codes=True)
ax=sns.barplot(x='Club',y='Assists',data=Assists_by_clubs.sort_values(by="Assists"),palette='S
ax.set_xlabel("Club",fontsize=30)
ax.set_ylabel("Assists",fontsize=20)
plt.xticks(rotation=75)
plt.rcParams["figure.figsize"]=(20,8)
plt.title('Plot of Clubs VS Total Assists',fontsize=20)
```

Out[115]:

```
Text(0.5, 1.0, 'Plot of Clubs VS Total Assists')
```



In [117]:

```
#Top Ten Assists
Top_ten_assists=epl_df[['Name','Club','Assists','Matches']].nlargest(n=10,columns='Assists')
Top_ten_assists
```

Out[117]:

	Name	Club	Assists	Matches
162	Harry Kane	Tottenham Hotspur	14	35
34	Kevin De Bruyne	Manchester City	12	25
51	Bruno Fernandes	Manchester United	12	37
161	Son Heung-min	Tottenham Hotspur	10	37
273	Jack Grealish	Aston Villa	10	26
54	Marcus Rashford	Manchester United	9	37

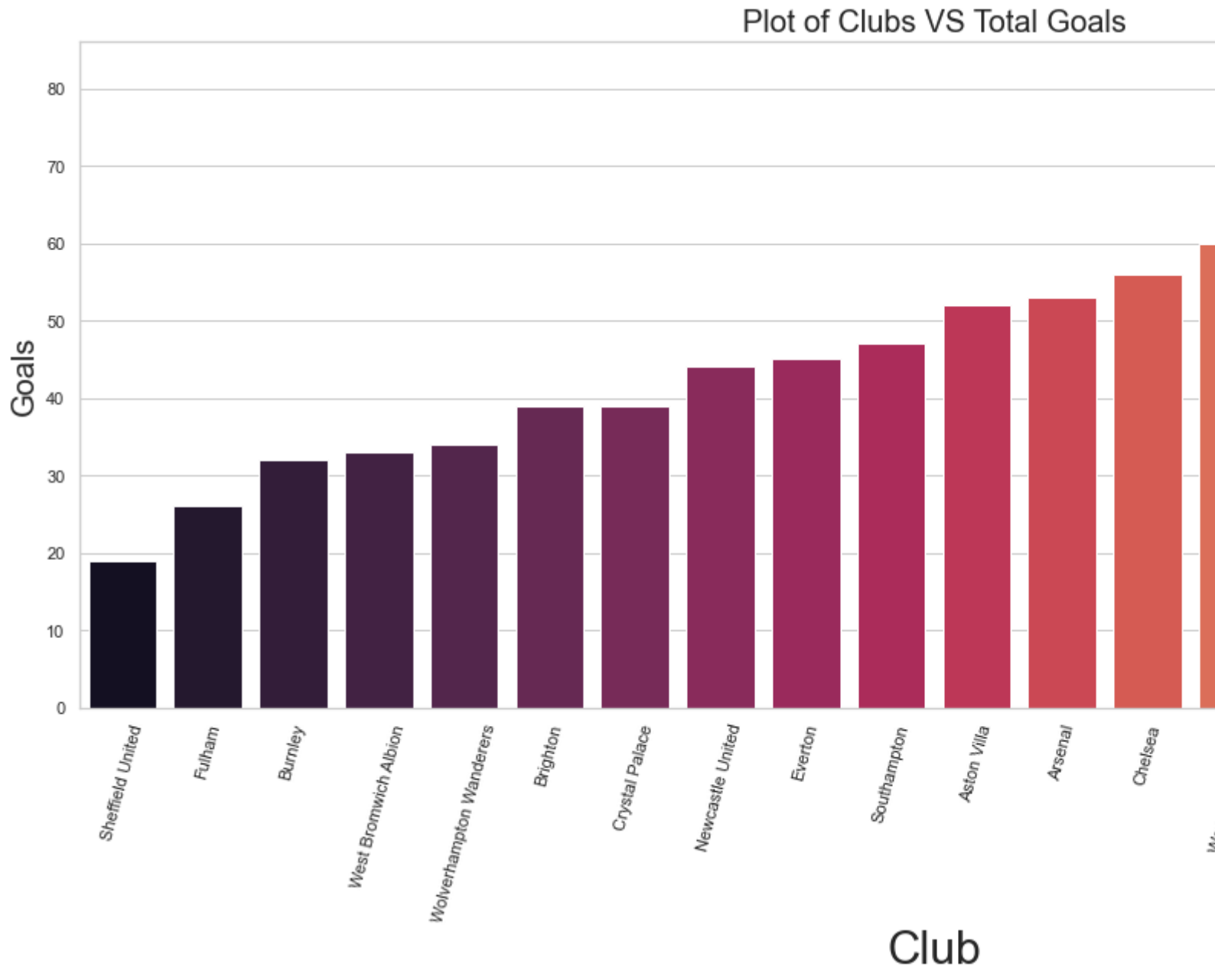
	Name	Club	Assists	Matches
110	Jamie Vardy	Leicester City	9	34
220	Raphael Dias Belloli	Leeds United	9	30
2	Timo Werner	Chelsea	8	35
136	Aaron Cresswell	West Ham United	8	36

In [120]:

```
Goals_by_clubs = pd.DataFrame(epl_df.groupby('Club',as_index=False)['Goals'].sum())
sns.set_theme(style="whitegrid",color_codes=True)
ax=sns.barplot(x='Club',y='Goals',data=Goals_by_clubs.sort_values(by="Goals"),palette='rocket'
ax.set_xlabel("Club",fontsize=30)
ax.set_ylabel("Goals",fontsize=20)
plt.xticks(rotation=75)
plt.rcParams["figure.figsize"]=(20,8)
plt.title('Plot of Clubs VS Total Goals',fontsize=20)
```

Out[120]:

```
Text(0.5, 1.0, 'Plot of Clubs VS Total Goals')
```



In [121]:

```
#Most goals scored  by player
Top_ten_Goals=epl_df[['Name','Club','Goals','Matches']].nlargest(n=10,columns='Goals')
Top_ten_Goals
```

Out[121]:

	Name	Club	Goals	Matches
162	Harry Kane	Tottenham Hotspur	23	35
81	Mohamed Salah	Liverpool FC	22	37
51	Bruno Fernandes	Manchester United	18	37
161	Son Heung-min	Tottenham Hotspur	17	37
214	Patrick Bamford	Leeds United	17	38
237	Dominic Calvert-Lewin	Everton	16	33
110	Jamie Vardy	Leicester City	15	34
267	Ollie Watkins	Aston Villa	14	37
33	?lkay G□ndo?an	Manchester City	13	28
191	Alexandre Lacazette	Arsenal	13	31

In [122]:

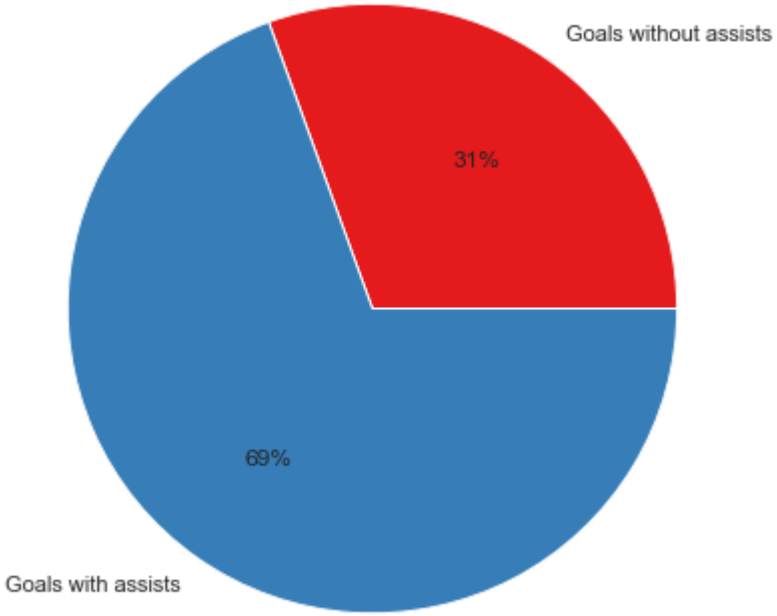
```
#Goal per match
top_10_goals_per_match=epl_df[['Name','GoalsPerMatch','Goals','Matches']].nlargest(n=10,column
top_10_goals_per_match
```

Out[122]:

	Name	GoalsPerMatch	Goals	Matches
162	Harry Kane	0.657143	23	35
81	Mohamed Salah	0.594595	22	37
307	Joe Willock	0.571429	8	14
145	Jesse Lingard	0.562500	9	16
175	Gareth Bale	0.550000	11	20
74	Anthony Elanga	0.500000	1	2
51	Bruno Fernandes	0.486486	18	37
237	Dominic Calvert-Lewin	0.484848	16	33
120	Kelechi Iheanacho	0.480000	12	25
92	Diogo Jota	0.473684	9	19

In [125]:

```
#Pie Chart -Goals with Assists and without Assists
plt.figure(figsize=(14,7))
assists=epl_df['Assists'].sum()
data=[Total_Goals - assists,assists]
labels=['Goals without assists','Goals with assists']
color=sns.color_palette('Set1')
plt.pie(data,labels=labels,colors=color,autopct='%.0f%%')
plt.show()
```



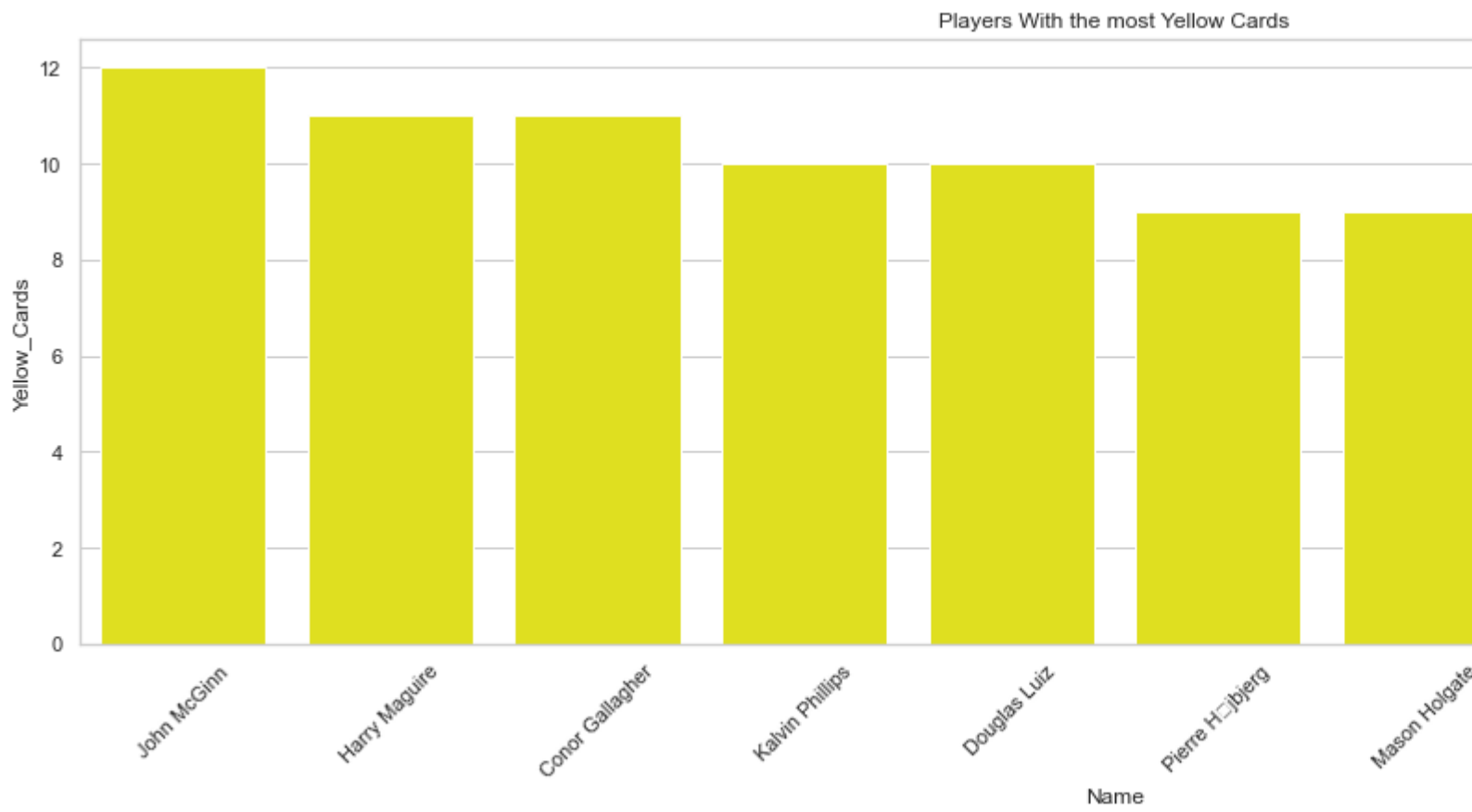
In [133]:

```
#Top 10 players with most yellow cards
epl_yellow=epl_df.sort_values(by='Yellow_Cards',ascending=False)[:10]
plt.figure(figsize=(20,6))
plt.title("Players With the most Yellow Cards")
c=sns.barplot(x=epl_yellow['Name'],y=epl_yellow['Yellow_Cards'],label='Player',color='yellow')
c.set_xticklabels(c.get_xticklabels(),rotation=45)
c
```

Out[133]:

```
<AxesSubplot:title={'center':'Players With the most Yellow Cards'}, xlabel='Name', ylabel='Yell

C:\Users\roshn\anaconda3\lib\site-packages\matplotlib\backends\backend_agg.py:238: RuntimeWarni
  font.set_text(s, 0.0, flags=flags)
C:\Users\roshn\anaconda3\lib\site-packages\matplotlib\backends\backend_agg.py:201: RuntimeWarni
  font.set_text(s, 0, flags=flags)
```


Players With the most Yellow Cards
Yellow_Cards
0
2
4
6
8
10
12
John McGinn
Harry Maguire
Conor Gallagher
Kalvin Phillips
Douglas Luiz
Pierre H□jbjerg
Mason Holgate
Name

In [ ]:

In [ ]:

In [ ]: