In [94]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [ ]:
```
athletes = pd.read_csv('C:/Users/amit bhardwaj/Desktop/Data Analysis Projects/olympic Analysis/Data set/athlete_events.cs
```

In [6]:
```
regions = pd.read_csv('C:/Users/amit bhardwaj/Desktop/Data Analysis Projects/olympic Analysis/Data set/noc_regions.csv')
```

In [9]:
```
athletes.head()
```

Out[9]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN

In [10]:
```
regions.head()
```

Out[10]:

	NOC	region	notes

	NOC	region	notes
0	AFG	Afghanistan	NaN
1	AHO	Curacao	Netherlands Antilles
2	ALB	Albania	NaN
3	ALG	Algeria	NaN
4	AND	Andorra	NaN

In [23]:
```
athletes_df = athletes.merge(regions, how = 'left', on = 'NOC')
```

In [24]:
```
athletes_df.head()
```

Out[24]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	r
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN	
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN	
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN	Der
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold	Der
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN	Nether

In [25]:
```
athletes_df.shape
```

Out[25]: (271116, 17)

In [31]:
```
athletes_df.rename(columns={'region':'edgar'}, inplace=True)
```

In [32]:
```
athletes_df.head()
```

Out[32]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN	
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN	
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN	Der
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold	Der
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN	Nether

In [37]:
```
athletes_df.rename(columns={'edgar':'Regions', 'notes': 'Notes'}, inplace=True)
```

In [38]:
```
athletes_df.head()
```

Out[38]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	Re
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN	

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	Re
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN	
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN	Der
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold	Der
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN	Nether

In [39]:
```
athletes_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 271116 entries, 0 to 271115
Data columns (total 17 columns):
 #   Column   Non-Null Count   Dtype  
---  ------   --------------   -----  
 0   ID       271116 non-null  int64  
 1   Name     271116 non-null  object 
 2   Sex      271116 non-null  object 
 3   Age      261642 non-null  float64
 4   Height   210945 non-null  float64
 5   Weight   208241 non-null  float64
 6   Team     271116 non-null  object 
 7   NOC      271116 non-null  object 
 8   Games    271116 non-null  object 
 9   Year     271116 non-null  int64  
 10  Season   271116 non-null  object 
 11  City     271116 non-null  object 
 12  Sport    271116 non-null  object 
 13  Event    271116 non-null  object 
 14  Medal    39783 non-null   object 
 15  Regions  270746 non-null  object 
 16  Notes    5039 non-null    object 
```

```
dtypes: float64(3), int64(2), object(12)
memory usage: 37.2+ MB
```

In [41]:
```
athletes_df.loc[0:4, ['Name', 'Age', 'Sex', 'NOC']]
```

Out[41]:

	Name	Age	Sex	NOC
0	A Dijiang	24.0	M	CHN
1	A Lamusi	23.0	M	CHN
2	Gunnar Nielsen Aaby	24.0	M	DEN
3	Edgar Lindenau Aabye	34.0	M	DEN
4	Christine Jacoba Aaftink	21.0	F	NED

In [42]:
```
athletes_df.describe()
```

Out[42]:

	ID	Age	Height	Weight	Year
count	271116.000000	261642.000000	210945.000000	208241.000000	271116.000000
mean	68248.954396	25.556898	175.338970	70.702393	1978.378480
std	39022.286345	6.393561	10.518462	14.348020	29.877632
min	1.000000	10.000000	127.000000	25.000000	1896.000000
25%	34643.000000	21.000000	168.000000	60.000000	1960.000000
50%	68205.000000	24.000000	175.000000	70.000000	1988.000000
75%	102097.250000	28.000000	183.000000	79.000000	2002.000000
max	135571.000000	97.000000	226.000000	214.000000	2016.000000

In [46]:
```
missing_values = athletes_df.isna()
missing_columns = missing_values.any()
missing_columns
```

Out[46]:
```
ID          False
Name        False
Sex         False
```

```
Age         True
Height      True
Weight      True
Team       False
NOC        False
Games      False
Year       False
Season     False
City       False
Sport      False
Event      False
Medal       True
Regions     True
Notes       True
dtype: bool
```

In [47]:
```
athletes_df.isnull().sum()
```

Out[47]:
```
ID             0
Name           0
Sex            0
Age         9474
Height     60171
Weight     62875
Team           0
NOC            0
Games          0
Year           0
Season         0
City           0
Sport          0
Event          0
Medal     231333
Regions      370
Notes     266077
dtype: int64
```

In [73]:
```
athletes_df.columns[athletes_df.isnull().any()].tolist()
```

Out[73]:
```
['Age', 'Height', 'Weight', 'Medal', 'Regions', 'Notes']
```

In [ ]:
```

```

In [91]:
```
athletes_df.query('Team == "India"'). head(5)
```

Out[91]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	Regions	Notes
505	281	S. Abdul Hamid	M	NaN	NaN	NaN	India	IND	1928 Summer	1928	Summer	Amsterdam	Athletics	Athletics Men's 110 metres Hurdles	NaN	India	NaN
506	281	S. Abdul Hamid	M	NaN	NaN	NaN	India	IND	1928 Summer	1928	Summer	Amsterdam	Athletics	Athletics Men's 400 metres Hurdles	NaN	India	NaN
895	512	Shiny Kurisingal Abraham-Wilson	F	19.0	167.0	53.0	India	IND	1984 Summer	1984	Summer	Los Angeles	Athletics	Athletics Women's 800 metres	NaN	India	NaN
896	512	Shiny Kurisingal Abraham-Wilson	F	19.0	167.0	53.0	India	IND	1984 Summer	1984	Summer	Los Angeles	Athletics	Athletics Women's 4 x 400 metres Relay	NaN	India	NaN
897	512	Shiny Kurisingal Abraham-Wilson	F	23.0	167.0	53.0	India	IND	1988 Summer	1988	Summer	Seoul	Athletics	Athletics Women's 800 metres	NaN	India	NaN

In [13]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [14]:
```
athletes = pd.read_csv('C:/Users/amit bhardwaj/Desktop/Data Analysis Projects/olympic Analysis/Data set/athlete_events.cs
```

In [15]:
```
regions = pd.read_csv('C:/Users/amit bhardwaj/Desktop/Data Analysis Projects/olympic Analysis/Data set/noc_regions.csv')
```

In [16]:
```
athletes_df = athletes.merge(regions, how = 'left', on = 'NOC')
```

In [17]:
```
athletes_df.head()
```

Out[17]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	r
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN	
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NaN	
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN	Der
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold	Der
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NaN	Nether

In [18]:
```
athletes_df.query('Team == "Japan"'). head(5)
```

Out[18]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	region	notes
625	362	Isao Ko Abe	M	24.0	177.0	75.0	Japan	JPN	1936 Summer	1936	Summer	Berlin	Athletics	Athletics Men's Hammer Throw	NaN	Japan	NaN
629	363	Kazumi Abe	M	28.0	178.0	67.0	Japan	JPN	1976 Winter	1976	Winter	Innsbruck	Bobsleigh	Bobsleigh Men's Four	NaN	Japan	NaN

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	region	notes
630	364	Kazuo Abe	M	25.0	166.0	69.0	Japan	JPN	1960 Summer	1960	Summer	Roma	Wrestling	Wrestling Men's Lightweight, Freestyle	NaN	Japan	NaN
631	365	Kinya Abe	M	23.0	168.0	68.0	Japan	JPN	1992 Summer	1992	Summer	Barcelona	Fencing	Fencing Men's Foil, Individual	NaN	Japan	NaN
632	366	Kiyoshi Abe	M	25.0	167.0	62.0	Japan	JPN	1972 Summer	1972	Summer	Munich	Wrestling	Wrestling Men's Featherweight, Freestyle	NaN	Japan	NaN

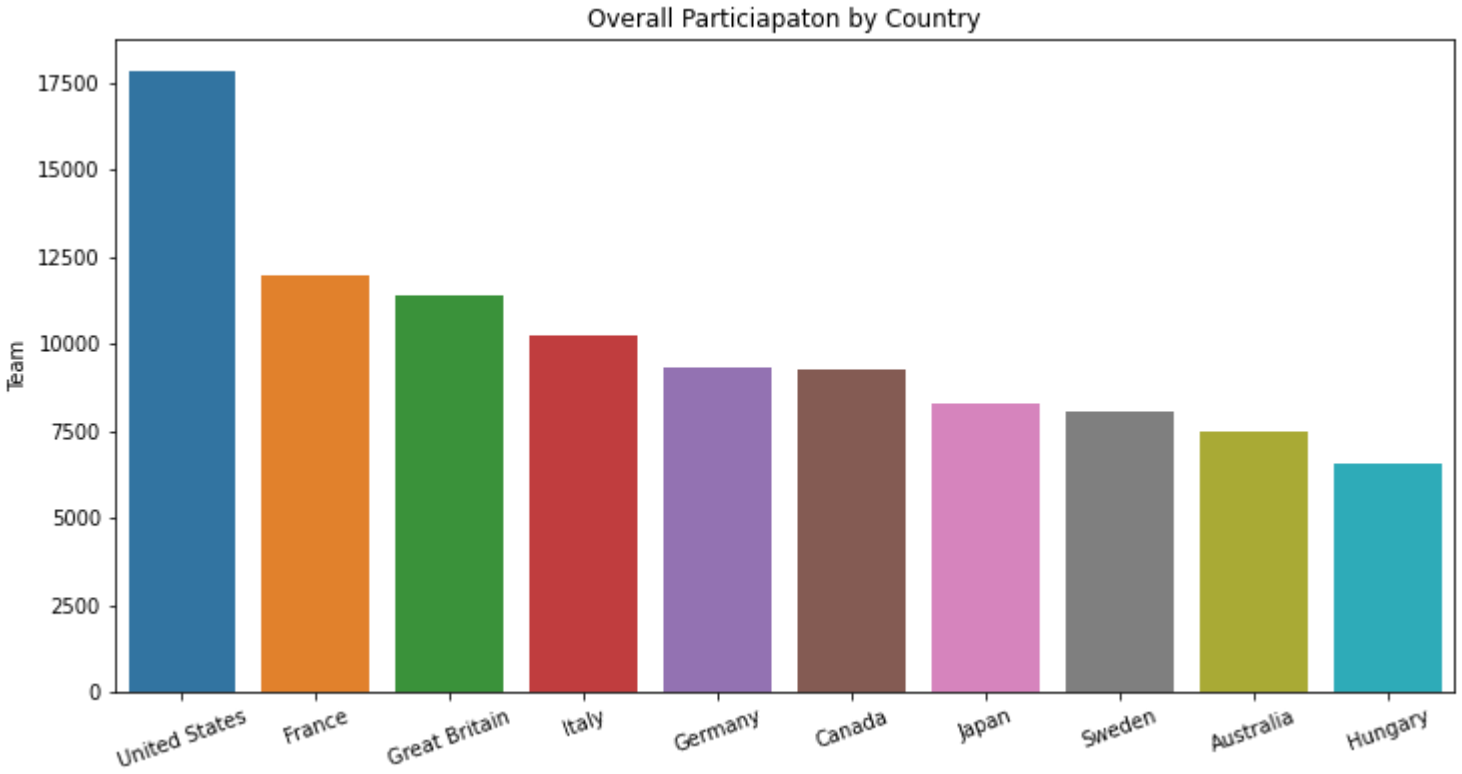
In [60]:
```
top_10_countries = athletes_df. Team.value_counts() .sort_values (ascending=False) .head(10)
```

In [61]:
```
top_10_countries
```

Out[61]:
```
United States    17847
France           11988
Great Britain    11404
Italy            10260
Germany           9326
Canada            9279
Japan             8289
Sweden            8052
Australia         7513
Hungary           6547
Name: Team, dtype: int64
```

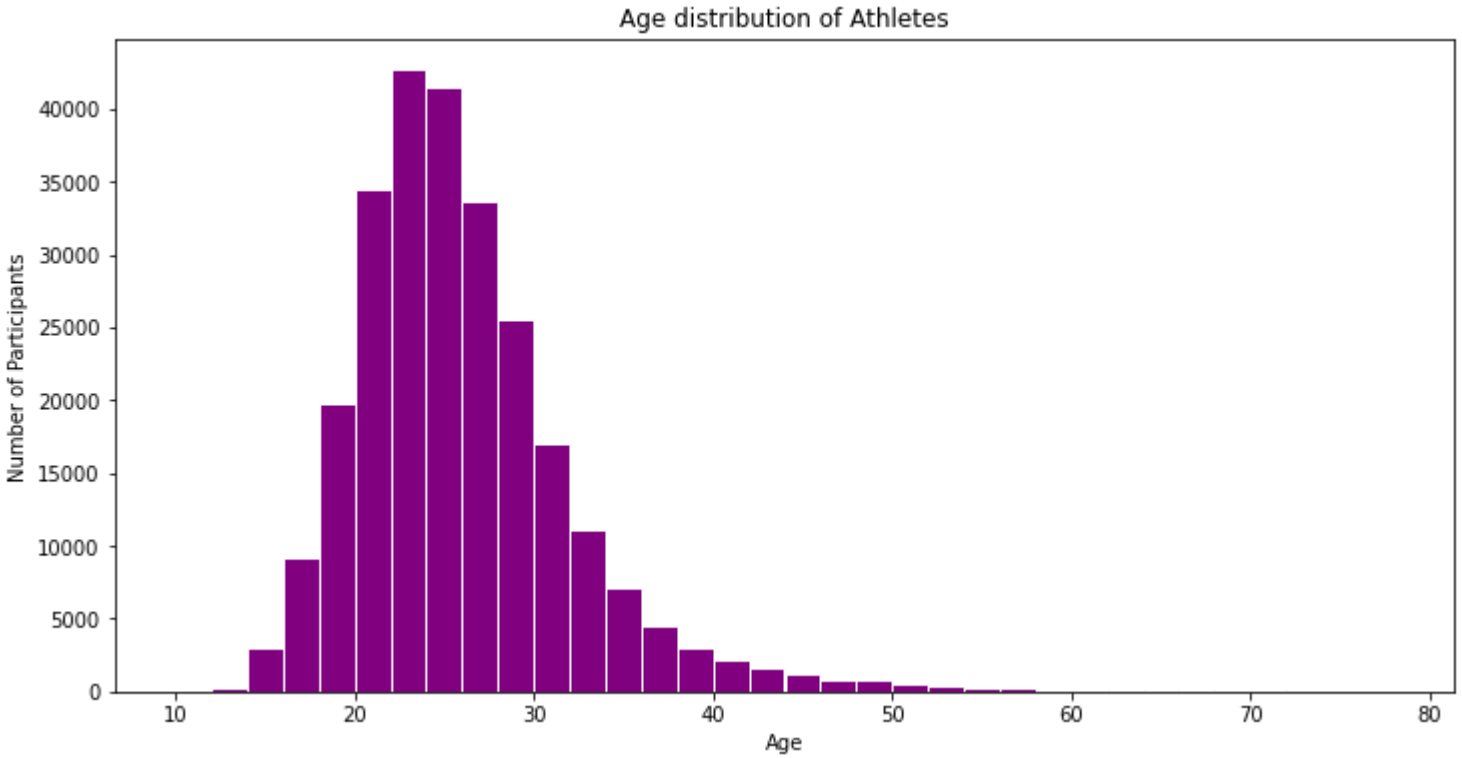
In [82]:
```
plt.figure(figsize=(12,6))
plt.xticks(rotation=20)
plt.title ('Overall Particiapaton by Country')
sns.barplot(x=top_10_countries.index, y=top_10_countries)
```

Out[82]:
```
<AxesSubplot:title={'center':'Overall Particiapaton by Country'}, ylabel='Team'>
```



In [96]:

```
plt.figure(figsize=(12,6))
plt.title("Age distribution of Athletes")
plt.xlabel( 'Age')
plt.ylabel( 'Number of Participants')
plt.hist (athletes_df.Age, bins=np.arange(10,80,2), color="purple", edgecolor= 'white');
```



In [98]:
```
winter_sports = athletes_df[athletes_df.Season =='Winter'].Sport.unique()
winter_sports
```

Out[98]:
```
array(['Speed Skating', 'Cross Country Skiing', 'Ice Hockey', 'Biathlon',
       'Alpine Skiing', 'Luge', 'Bobsleigh', 'Figure Skating',
       'Nordic Combined', 'Freestyle Skiing', 'Ski Jumping', 'Curling',
       'Snowboarding', 'Short Track Speed Skating', 'Skeleton',
       'Military Ski Patrol', 'Alpinism'], dtype=object)
```

In [99]:
```
summer_sports = athletes_df[athletes_df.Season =='Summer'].Sport.unique()
summer_sports
```

Out[99]:
```
array(['Basketball', 'Judo', 'Football', 'Tug-Of-War', 'Athletics',
       'Swimming', 'Badminton', 'Sailing', 'Gymnastics',
       'Art Competitions', 'Handball', 'Weightlifting', 'Wrestling',
       'Water Polo', 'Hockey', 'Rowing', 'Fencing', 'Equestrianism',
       'Shooting', 'Boxing', 'Taekwondo', 'Cycling', 'Diving', 'Canoeing',
       'Tennis', 'Modern Pentathlon', 'Golf', 'Softball', 'Archery',
       'Volleyball', 'Synchronized Swimming', 'Table Tennis', 'Baseball',
```

```
'Rhythmic Gymnastics', 'Rugby Sevens', 'Trampolining',
'Beach Volleyball', 'Triathlon', 'Rugby', 'Lacrosse', 'Polo',
'Cricket', 'Ice Hockey', 'Racquets', 'Motorboating', 'Croquet',
'Figure Skating', 'Jeu De Paume', 'Roque', 'Basque Pelota',
'Alpinism', 'Aeronautics'], dtype=object)
```

In [ ]: