In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [6]:
```
# Load Dataset

athletes = pd.read_csv('C:/Users/ezhil/OneDrive/Desktop/Olympics Dataset/athlete_events.csv')
regions = pd.read_csv('C:/Users/ezhil/OneDrive/Desktop/Olympics Dataset/noc_regions.csv')
```

In [4]:
```
athletes.head()
```

Out[4]:

	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 [7]:
```
regions.head()
```

Out[7]:

	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 [9]:
```
# Join the dataframes

athletes_df = athletes.merge(regions, how = 'left', on = 'NOC')
athletes_df.head()
```

Out[9]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	region	notes
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN	China	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	China	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN	Denmark	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	Denmark	NaN
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	Netherlands	NaN

In [10]:
```
athletes_df.shape
```

Out[10]: (271116, 17)

In [12]:
```
# Column names consistent

athletes_df.rename(columns={'region':'Region', 'notes':'Notes'}, inplace=True)
```

In [13]:
```
athletes_df.head()
```

Out[13]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	Region	Notes
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN	China	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	China	NaN
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN	Denmark	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	Denmark	NaN
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	Netherlands	NaN

In [14]:
```
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  Region  270746 non-null  object
 16  Notes   5039 non-null    object
dtypes: float64(3), int64(2), object(12)
memory usage: 37.2+ MB
```

In [15]:
```
athletes_df.describe()
```

Out[15]:

	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 [16]:
```
# Check null values
nan_values = athletes_df.isna()
nan_columns = nan_values.any()
nan_columns
```

Out[16]:
```
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
Region     True
Notes      True
dtype: bool
```

In [17]:
```
athletes_df.isnull().sum()
```

Out[17]:
```
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
Region       370
Notes     266077
dtype: int64
```

In [18]:
```
# India details

athletes_df.query('Team == "India"').head(5)
```

Out[18]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	Region	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 [19]:
```
# Japan details

athletes_df.query('Team == "Japan"').head(5)
```

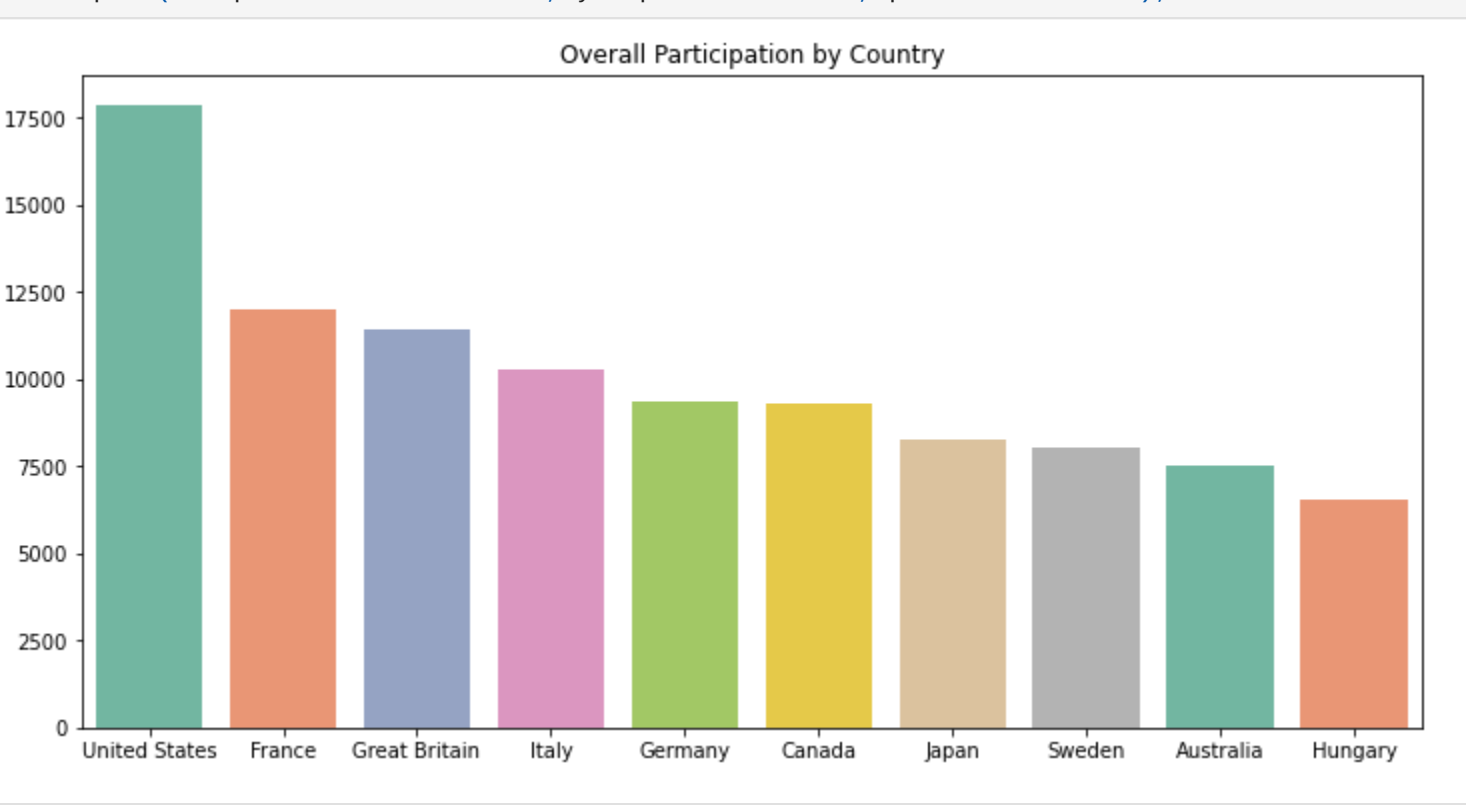
Out[19]:

	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
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

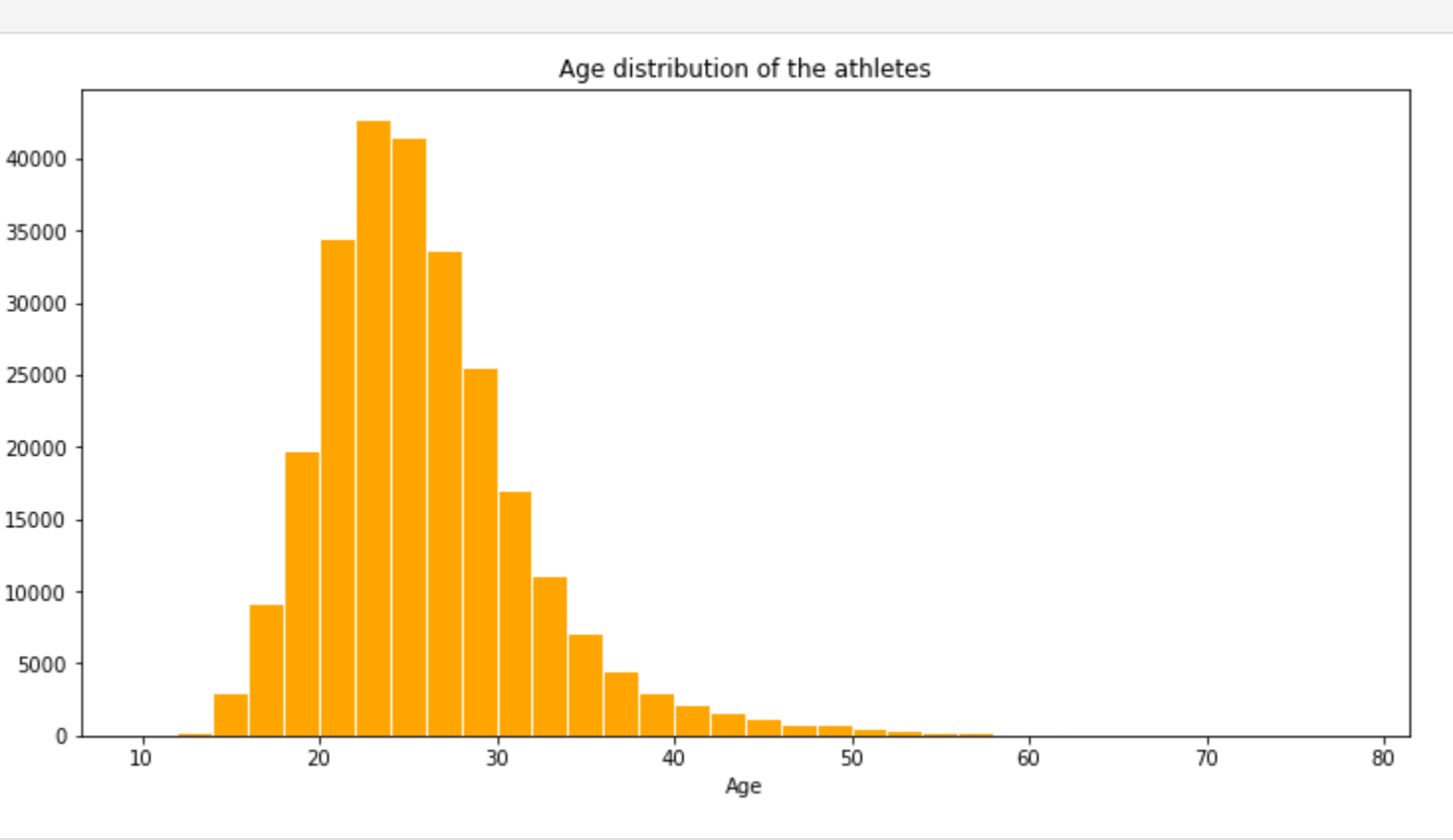
In [20]:
```
# Top countries participating

top_10_countries = athletes_df.Team.value_counts().sort_values(ascending=False).head(10)
top_10_countries
```

Out[20]:
```
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 [22]:
```
# Plot for the top 10 countries

plt.figure(figsize=(12,6))
plt.title("Overall Participation by Country")
sns.barplot(x=top_10_countries.index, y=top_10_countries, palette = 'Set2');
```



In [28]:
```
# Age Distribution of the participations

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



In [29]:
```
# Winter olympics sports

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

Out[29]:
```
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)
```

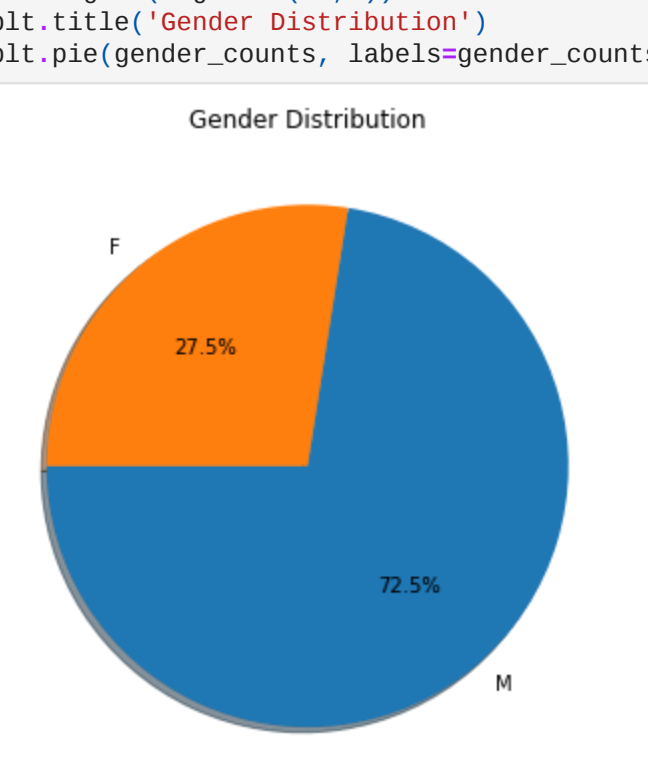
In [30]:
```
# Summer olympics sports

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

Out[30]:
```
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 [31]:
```
# Male and Female participants

gender_counts = athletes_df.Sex.value_counts()
gender_counts
```

Out[31]:
```
M    196594
F     74522
Name: Sex, dtype: int64
```

In [32]:
```
# Pie plot for male and female athletes

plt.figure(figsize=(12,6))
plt.title('Gender Distribution')
plt.pie(gender_counts, labels=gender_counts.index, autopct='%1.1f%%', startangle=180, shadow=True);
```



In [33]:
```
# Total medals

athletes_df.Medal.value_counts()
```

Out[33]:
```
Gold      13372
Bronze    13295
Silver    13116
Name: Medal, dtype: int64
```

In [35]:
```
# Total number of female athletes in each olympics

female_participants = athletes_df[(athletes_df.Sex=='F') & (athletes_df.Season == 'Summer')][['Sex', 'Year']]
female_participants = female_participants.groupby('Year').count().reset_index()
female_participants.tail()
```

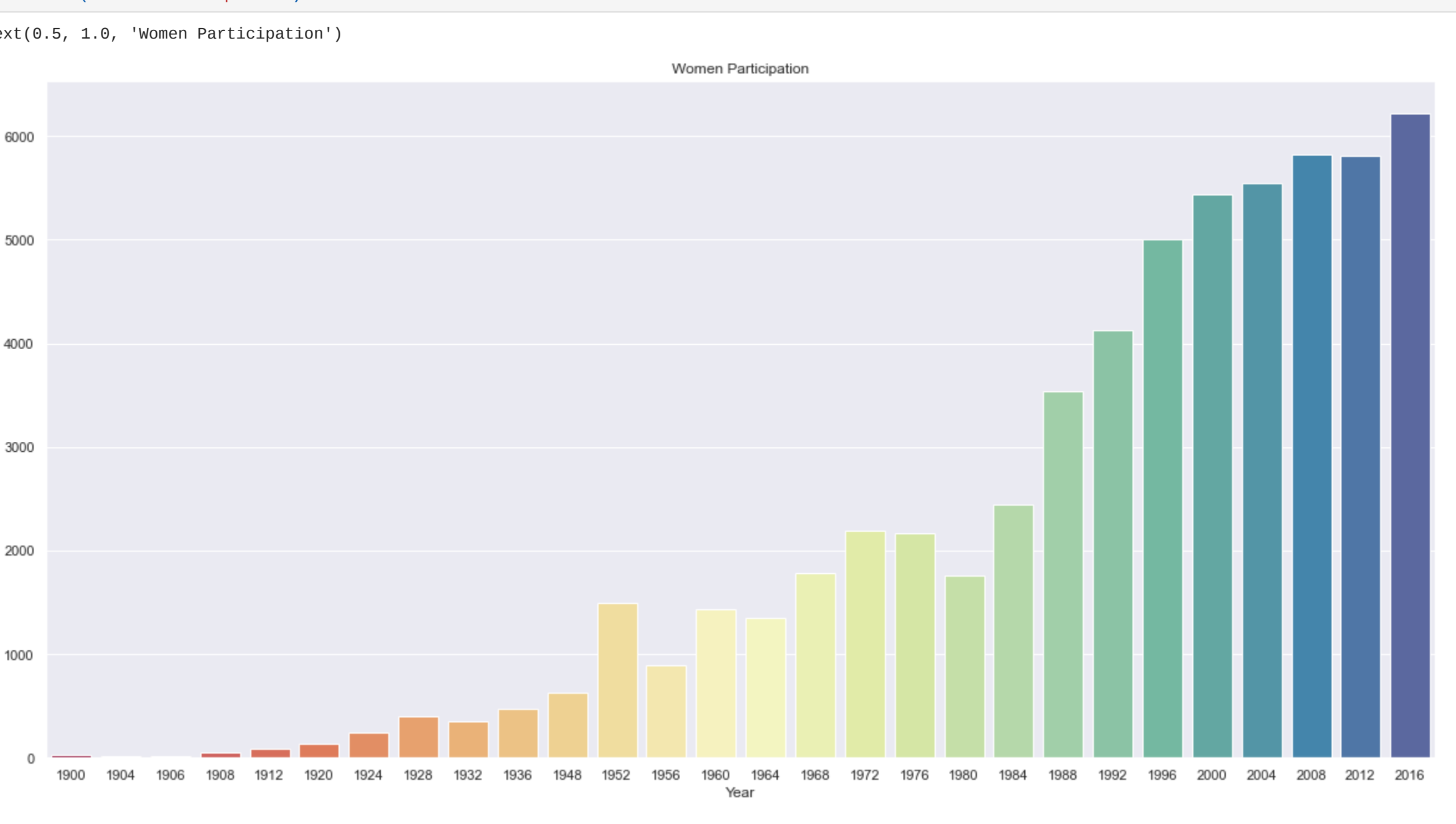
Out[35]:

	Year	Sex
23	2000	5431
24	2004	5546
25	2008	5816
26	2012	5815
27	2016	6223

In [36]:
```
womenOlympics = athletes_df[(athletes_df.Sex == 'F') & (athletes_df.Season == 'Summer')]
```

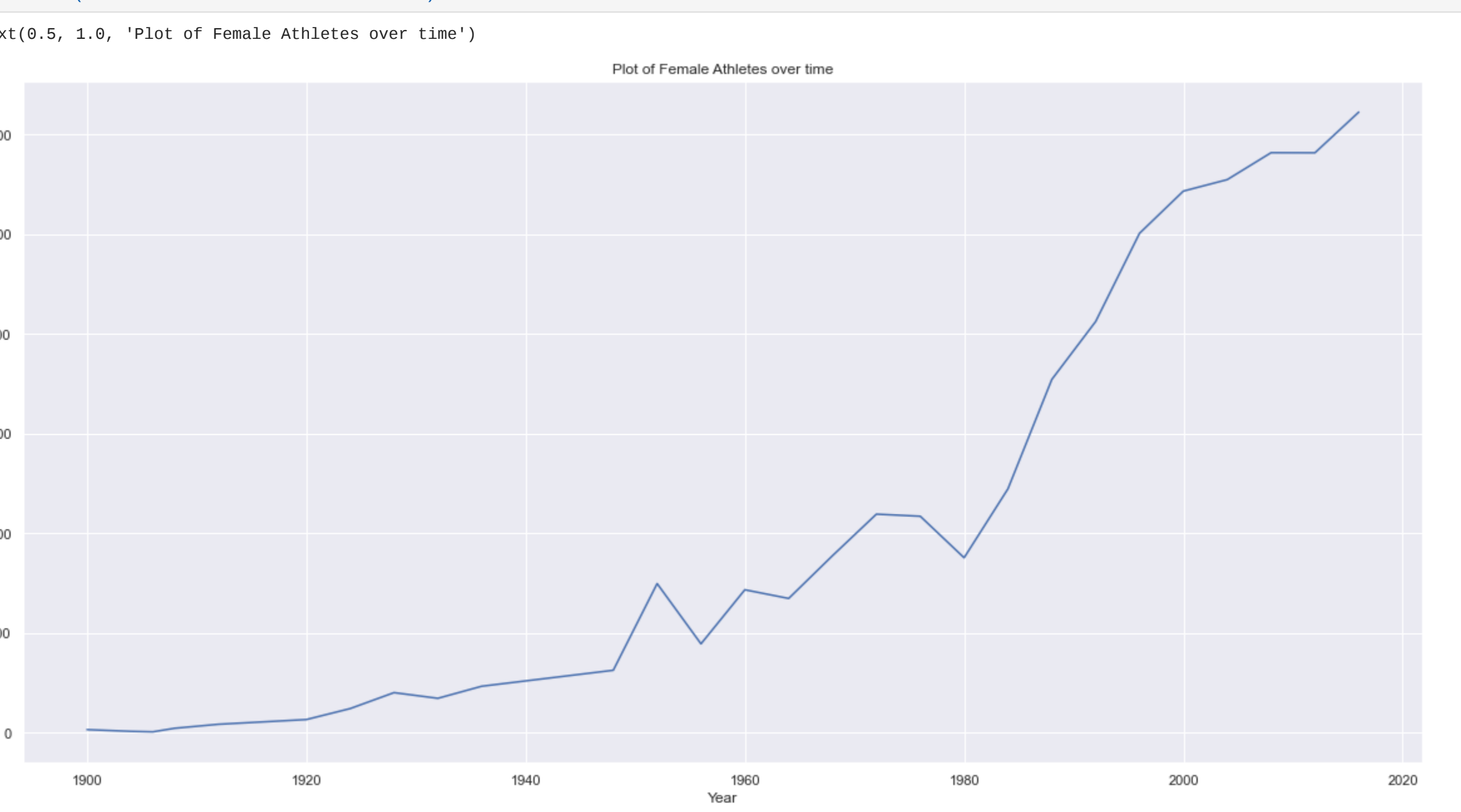
In [37]:
```
sns.set(style="darkgrid")
plt.figure(figsize=(20, 10))
sns.countplot(x='Year', data=womenOlympics, palette ="Spectral")
plt.title('Women Participation')
```

Out[37]: Text(0.5, 1.0, 'Women Participation')



In [38]:
```
part = womenOlympics.groupby('Year')['Sex'].value_counts()
plt.figure(figsize=(20,10))
part.loc[:,'F'].plot()
plt.title('Plot of Female Athletes over time')
```

Out[38]: Text(0.5, 1.0, 'Plot of Female Athletes over time')



In [52]:
```
#Gold medal athletes

goldMedals = athletes_df[(athletes_df.Medal == 'Gold')]
goldMedals.head()
```

Out[52]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	Region	Notes
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	Denmark	NaN
42	17	Paavo Johannes Aaltonen	M	28.0	175.0	64.0	Finland	FIN	1948 Summer	1948	Summer	London	Gymnastics	Gymnastics Men's Team All-Around	Gold	Finland	NaN
44	17	Paavo Johannes Aaltonen	M	28.0	175.0	64.0	Finland	FIN	1948 Summer	1948	Summer	London	Gymnastics	Gymnastics Men's Horse Vault	Gold	Finland	NaN
48	17	Paavo Johannes Aaltonen	M	28.0	175.0	64.0	Finland	FIN	1948 Summer	1948	Summer	London	Gymnastics	Gymnastics Men's Pommelled Horse	Gold	Finland	NaN
60	20	Kjetil Andr Aamodt	M	20.0	176.0	85.0	Norway	NOR	1992 Winter	1992	Winter	Albertville	Alpine Skiing	Alpine Skiing Men's Super G	Gold	Norway	NaN

In [53]:
```
#  take only the values that are different from NaN

goldMedals = goldMedals[np.isfinite(goldMedals['Age'])]
```

In [54]:
```
# Gold beyond 60

goldMedals['ID'][goldMedals['Age'] > 60].count()
```

Out[54]: 6

In [55]:
```
sporting_event = goldMedals['Sport'][goldMedals['Age']>60]
sporting_event
```

Out[55]:
```
104003    Art Competitions
105199               Roque
190952             Archery
226374             Archery
233390            Shooting
261102             Archery
Name: Sport, dtype: object
```

In [58]:
```
# Plot for sporting_event

plt.figure(figsize=(10,5))
plt.tight_layout()
sns.countplot(x=sporting_event)
plt.title('Gold Medals for athletes over 60 years')
```

Out[58]: Text(0.5, 1.0, 'Gold Medals for athletes over 60 years')

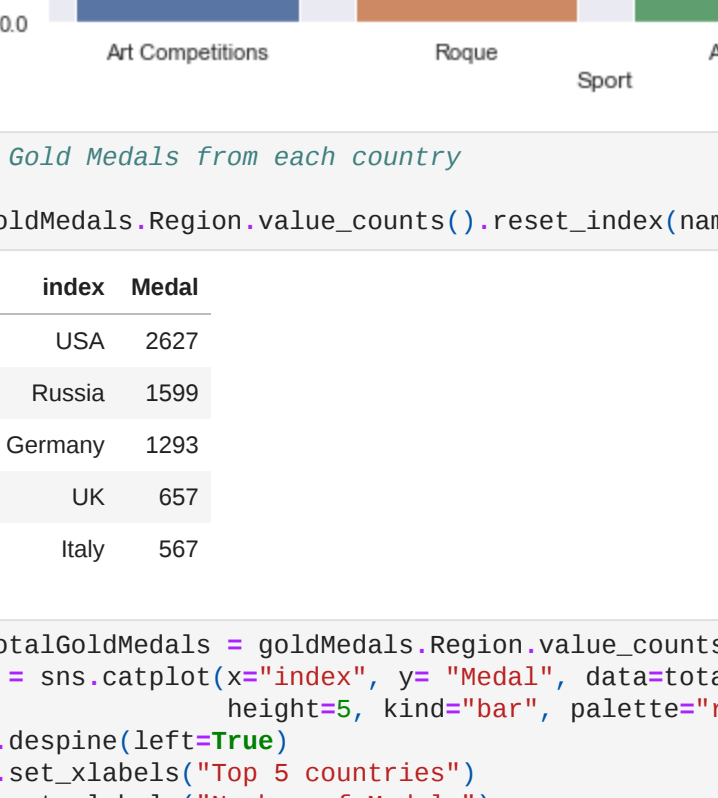
In [59]:
```
# Gold Medals from each country

goldMedals.Region.value_counts().reset_index(name='Medal').head(5)
```

Out[59]:

	index	Medal
0	USA	2627
1	Russia	1599
2	Germany	1293
3	UK	657
4	Italy	567

In [62]:
```
totalGoldMedals = goldMedals.Region.value_counts().reset_index(name='Medal').head(6)
g = sns.catplot(x="index", y= "Medal", data=totalGoldMedals,
                height=5, kind="bar", palette="rocket")
g.despine(left=True)
g.set_xlabels("Top 5 countries")
g.set_ylabels("Number of Medals")
plt.title('Gold Medals per Country')
```

Out[62]: Text(0.5, 1.0, 'Gold Medals per Country')

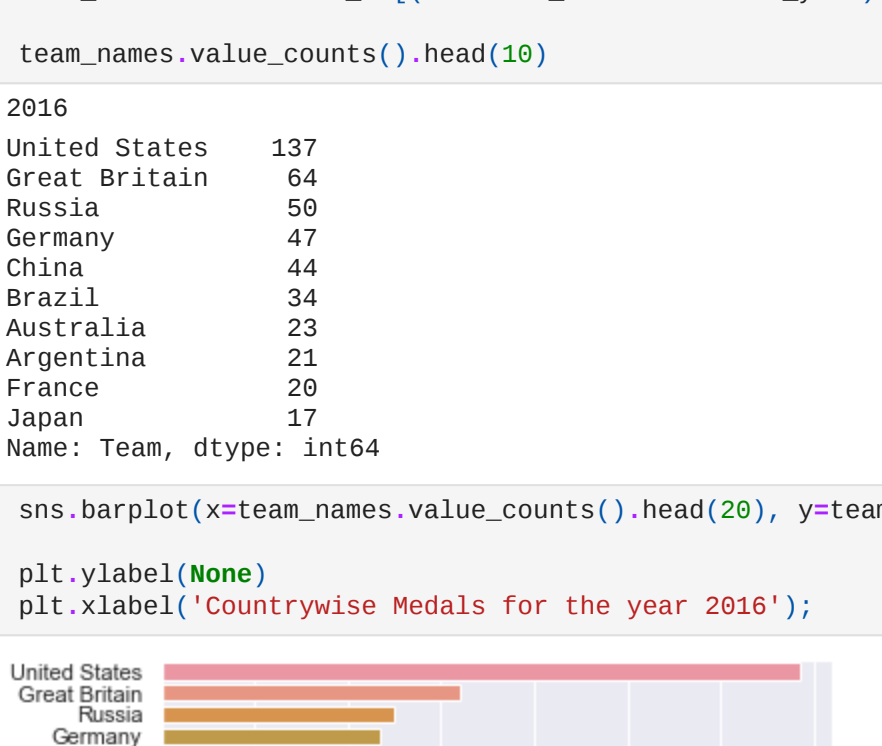


In [64]:
```
# Rio olympics

max_year = athletes_df.Year.max()
print(max_year)

team_names = athletes_df[(athletes_df.Year == max_year) & (athletes_df.Medal == 'Gold')].Team

team_names.value_counts().head(10)
```

```
2016
```

Out[64]:
```
United States    137
Great Britain     64
Russia            50
Germany           47
China             44
Brazil            34
Australia         23
Argentina         21
France            20
Japan             17
Name: Team, dtype: int64
```

In [65]:
```
sns.barplot(x=team_names.value_counts().head(20), y=team_names.value_counts().head(20).index)

plt.ylabel(None)
plt.xlabel('Countrywise Medals for the year 2016');
```



In [ ]: