In [18]:

```
#importing pandas
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [21]:

```
athletes =   pd.read_csv('athlete_events.csv')#reading csv using pandas
region   =   pd.read_csv('noc_regions.csv')
```

In [4]:

```
athletes.head()#showing head of dataframe
```

Out[4]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilitie
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPu
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPu
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPu
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPu
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPu

5 rows × 81 columns

In [22]:

```
region.head()
```

Out[22]:

	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 [24]:

```
#combining athlete and region data frame
athlete_merge = athletes.merge(region,how = 'left',on ='NOC')
```

In [26]:

```
athlete_merge.head()
```

Out[26]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barce
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	Lor
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwe
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	F
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Cal

In [28]:

```
#Checking the number of records
athlete_merge.shape
```

Out[28]:

```
(271116, 17)
```

In [32]:

```
#Renaming Column names
athlete_merge.rename(columns={'region':'Region','notes':'Notes'},inplace=Tr
```

In [33]:

```
athlete_merge.head()
```

Out[33]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barce
1	2	A Lamusi	M	23.0	170.0	60.0	China	CHN	2012 Summer	2012	Summer	Lor
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwe
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Summer	F
4	5	Christine Jacoba Aaftink	F	21.0	185.0	82.0	Netherlands	NED	1988 Winter	1988	Winter	Cal

Type *Markdown* and LaTeX: α^2

In [35]:

```
athlete_merge.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 [36]:

```
#statistical info using describe
athlete_merge.describe()
```

Out[36]:

	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 [42]:

```
# check which column  has null values(not needed)
null_value   = athlete_merge.isna()
null_columns = null_value.any()
null_columns
```

Out[42]:

```
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 [43]:

```
#calculating the number of null values on each column
athlete_merge.isnull().sum()
```

Out[43]:

```
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 [51]:

```
#Details of specific country
athlete_merge.query('Team == "India"').head()
```

Out[51]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City
505	281	S. Abdul Hamid	M	NaN	NaN	NaN	India	IND	1928 Summer	1928	Summer	Amsterdam
506	281	S. Abdul Hamid	M	NaN	NaN	NaN	India	IND	1928 Summer	1928	Summer	Amsterdam
895	512	Shiny Kurisingal Abraham-Wilson	F	19.0	167.0	53.0	India	IND	1984 Summer	1984	Summer	Los Angeles
896	512	Shiny Kurisingal Abraham-Wilson	F	19.0	167.0	53.0	India	IND	1984 Summer	1984	Summer	Los Angeles
897	512	Shiny Kurisingal Abraham-Wilson	F	23.0	167.0	53.0	India	IND	1988 Summer	1988	Summer	Seoul

In [59]:

```
#top 20 countries
top_20_countries = athlete_merge.Team.value_counts().sort_values(ascending=
```

In [60]:

```
top_20_countries
```

Out[60]:

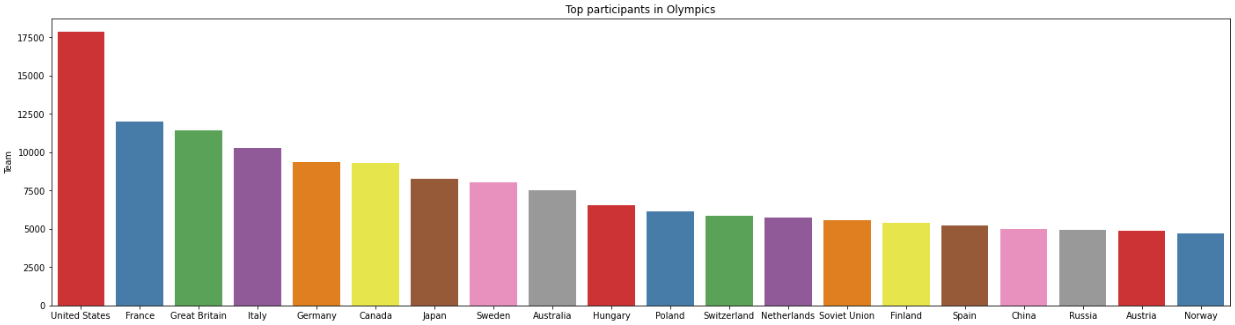
```
United States    17847
France           11988
Great Britain    11404
Italy            10260
Germany           9326
Canada            9279
Japan             8289
Sweden            8052
Australia         7513
Hungary           6547
Poland            6143
Switzerland       5844
Netherlands       5718
Soviet Union      5535
Finland           5379
Spain             5224
China             4975
Russia            4922
Austria           4866
Norway            4708
Name: Team, dtype: int64
```

In [65]:

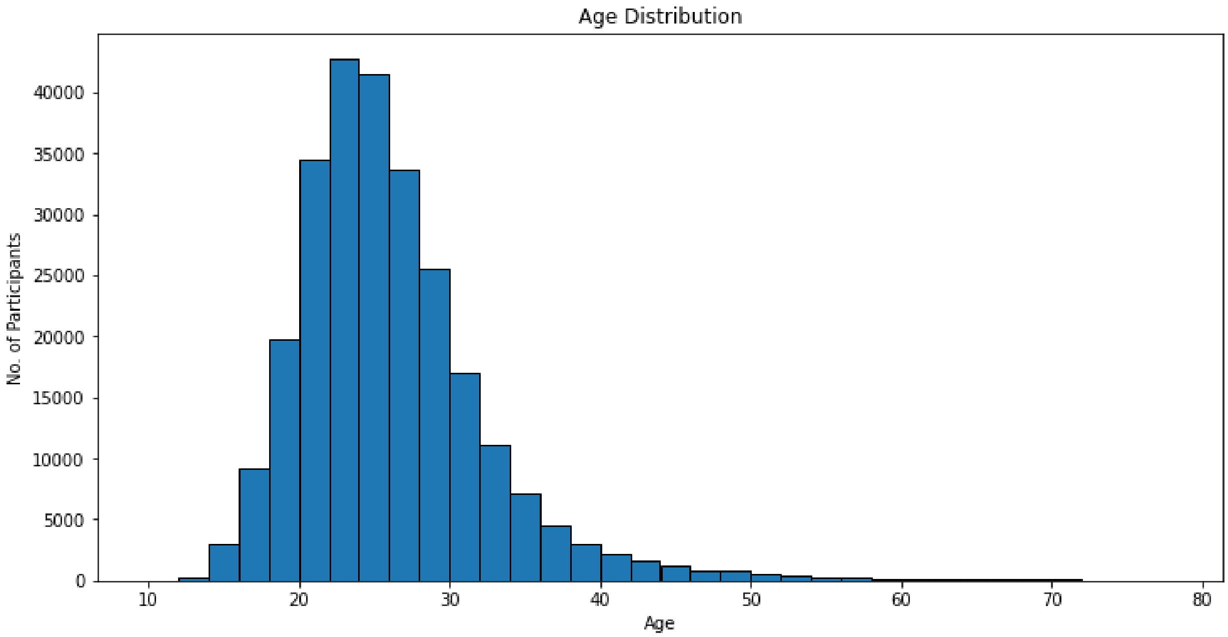
```
# Bar plot
plt.figure(figsize=(24,6))
plt.title('Top participants in Olympics')
sns.barplot(x=top_20_countries.index,y=top_20_countries,palette = 'Set1')
```

Out[65]:

```
<AxesSubplot:title={'center':'Top participants in Olympics'}, ylabel='Tea
m'>
```



In [88]:
```
#Age Distribution of athletes
plt.figure(figsize=(12,6))
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('No. of Participants')
plt.hist(athlete_merge.Age, bins = np.arange(10,80,2),edgecolor = 'Black');
```



In [93]:
```
#Gender Distribution
gender_count = athlete_merge.Sex.value_counts()
gender_count
```

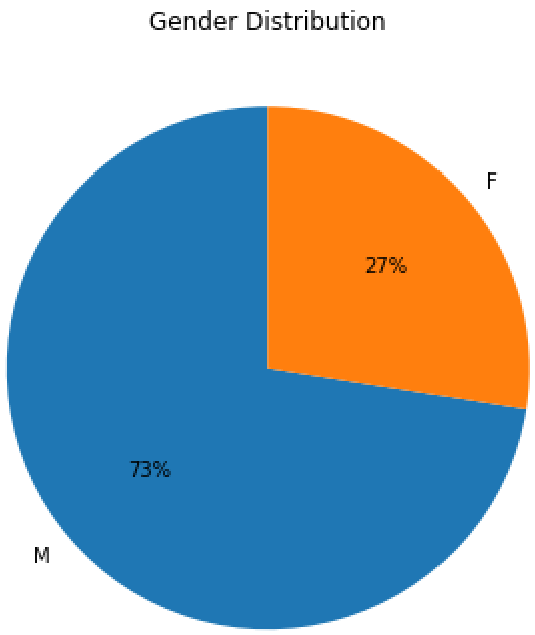
Out[93]:
```
M    196594
F     74522
Name: Sex, dtype: int64
```

In [104]:

```
#pie chart for gender distribution
plt.figure(figsize=(24,6))
plt.title('Gender Distribution')
plt.pie(gender_count,labels=gender_count.index,startangle = 90,autopct = '%
```

Out[104]:

```
([<matplotlib.patches.Wedge at 0x7fc0c6a0c040>,
  <matplotlib.patches.Wedge at 0x7fc0c6a1d730>],
 [Text(-0.8361576252945936, -0.7147310163003325, 'M'),
  Text(0.8361576922125369, 0.7147309380136029, 'F')],
 [Text(-0.4560859774334146, -0.38985328161836313, '73%'),
  Text(0.45608601393411097, 0.38985323891651064, '27%')])
```



In [107]:

```
#Total medal count
athlete_merge.Medal.value_counts()
```

Out[107]:

```
Gold      13372
Bronze    13295
Silver    13116
Name: Medal, dtype: int64
```

In [136]:

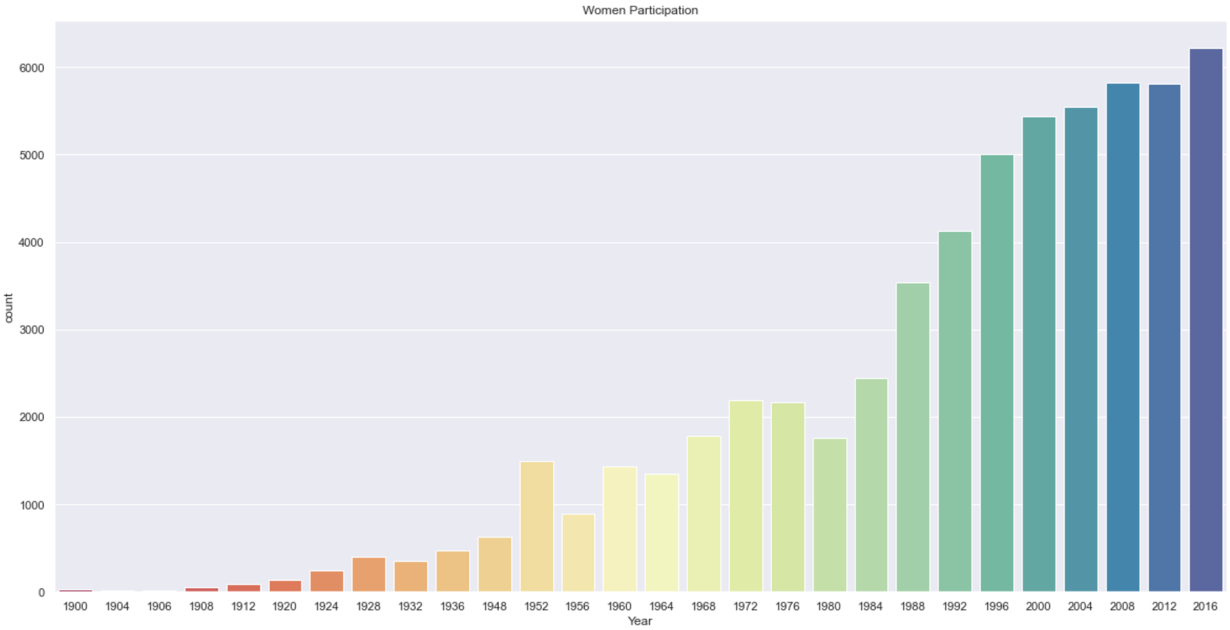
```
WomenInOlympics = athlete_merge[(athlete_merge.Sex=='F')&(athlete_merge.Sea
```

In [137]:

```
#women participation
sns.set(style="darkgrid")
plt.figure(figsize=(20,10))
sns.countplot(x='Year',data=WomenInOlympics,palette="Spectral")
plt.title('Women Participation')
```

Out[137]:

```
Text(0.5, 1.0, 'Women Participation')
```



In [128]:

```
#Female athletes in olympics
female_athlete= athlete_merge[(athlete_merge.Sex == 'F') & (athlete_merge.S
female_athlete=female_athlete.groupby('Year').count().reset_index()
female_athlete
```

Out[128]:

	Year	Sex
0	1924	17
1	1928	33
2	1932	22
3	1936	81
4	1948	133
5	1952	185
6	1956	246
7	1960	295
8	1964	404
9	1968	416
10	1972	415
11	1976	434
12	1980	430
13	1984	536
14	1988	680
15	1992	1054
16	1994	1105
17	1998	1384
18	2002	1582
19	2006	1757
20	2010	1847
21	2014	2023

In [147]:

```
#gold beyond 40
athlete_gold = athlete_merge[(athlete_merge.Sex == 'M') &  (athlete_merge.M
athlete_gold
```

Out[147]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Sea
3	4	Edgar Lindenau Aabye	M	34.0	NaN	NaN	Denmark/Sweden	DEN	1900 Summer	1900	Sum
42	17	Paavo Johannes Aaltonen	M	28.0	175.0	64.0	Finland	FIN	1948 Summer	1948	Sum
44	17	Paavo Johannes Aaltonen	M	28.0	175.0	64.0	Finland	FIN	1948 Summer	1948	Sum
48	17	Paavo Johannes Aaltonen	M	28.0	175.0	64.0	Finland	FIN	1948 Summer	1948	Sum
60	20	Kjetil Andr Aamodt	M	20.0	176.0	85.0	Norway	NOR	1992 Winter	1992	Wi
...	...	...	...	...	...	...	...	...	...	...	
270896	135474	Albert Hermann Zrner	M	18.0	NaN	NaN	Germany	GER	1908 Summer	1908	Sum
270917	135481	Jules Alexis "Louis" Zutter	M	30.0	NaN	NaN	Switzerland	SUI	1896 Summer	1896	Sum
270981	135503	Zurab Zviadauri	M	23.0	182.0	90.0	Georgia	GEO	2004 Summer	2004	Sum
271016	135523	Ronald Ferdinand "Ron" Zwerver	M	29.0	200.0	93.0	Netherlands	NED	1996 Summer	1996	Sum
271049	135545	Henk Jan Zwolle	M	31.0	197.0	93.0	Netherlands	NED	1996 Summer	1996	Sum

9625 rows × 17 columns

In [149]:

```
#gold medal above  the age of 40
gold_40 = athlete_gold[(athlete_gold.Age>=40)]
gold_40
```

Out[149]:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season
1755	974	Nils August Domingo Adlercreutz	M	45.0	NaN	NaN	Sweden	SWE	1912 Summer	1912	Summer
3306	1858	Fehaid Al-Deehani	M	49.0	178.0	95.0	Individual Olympic Athletes	IOA	2016 Summer	2016	Summer
3542	2025	Ahmed bin Hasher Al-Maktoum	M	40.0	175.0	67.0	United Arab Emirates	UAE	2004 Summer	2004	Summer
4784	2735	Sergey Gennadyevich Alifirenko	M	41.0	168.0	72.0	Russia	RUS	2000 Summer	2000	Summer
4878	2785	Alphonse Allaert	M	44.0	NaN	NaN	Belgium	BEL	1920 Summer	1920	Summer
...	...	...	...	...	...	...	...	...	...	...	...
261845	130999	Hans Gnter Winkler	M	46.0	174.0	72.0	West Germany	FRG	1972 Summer	1972	Summer
263201	131700	Frank Seymour Wright	M	41.0	174.0	NaN	United States	USA	1920 Summer	1920	Summer
266293	133226	Mahonri Mackintosh Young	M	54.0	NaN	NaN	United States	USA	1932 Summer	1932	Summer
267813	133986	Jzef Zapdzki	M	43.0	174.0	71.0	Poland	POL	1972 Summer	1972	Summer
269922	135045	Rbert Zimonyi	M	46.0	170.0	52.0	United States	USA	1964 Summer	1964	Summer

378 rows × 17 columns

In [ ]: