120 Years Olympic Data Analysis

```
# Importing Libraries

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

# Loading the Dataset

atheletes = pd.read_csv("C:\\Users\\PRANAY\\Downloads\\dataset\\
athlete_events.csv")
regions = pd.read_csv("C:\\Users\\PRANAY\\Downloads\\dataset\\
noc_regions.csv")

atheletes.head()
```

```
   ID                      Name Sex   Age  Height  Weight 
Team  \
0   1                 A Dijiang   M  24.0   180.0    80.0 
China
1   2                  A Lamusi   M  23.0   170.0    60.0 
China
2   3       Gunnar Nielsen Aaby   M  24.0     NaN     NaN 
Denmark
3   4      Edgar Lindenau Aabye   M  34.0     NaN     NaN 
Denmark/Sweden
4   5  Christine Jacoba Aaftink   F  21.0   185.0    82.0 
Netherlands

   NOC        Games  Year  Season       City          Sport  \
0  CHN  1992 Summer  1992  Summer  Barcelona     Basketball
1  CHN  2012 Summer  2012  Summer     London           Judo
2  DEN  1920 Summer  1920  Summer  Antwerpen       Football
3  DEN  1900 Summer  1900  Summer      Paris     Tug-Of-War
4  NED  1988 Winter  1988  Winter    Calgary  Speed Skating

                              Event Medal
0       Basketball Men's Basketball   NaN
1      Judo Men's Extra-Lightweight   NaN
2           Football Men's Football   NaN
3       Tug-Of-War Men's Tug-Of-War  Gold
4  Speed Skating Women's 500 metres   NaN
```

```
regions.head()
```

```
   NOC       region                    notes
0  AFG  Afghanistan                      NaN
```

```
1  AHO      Curacao  Netherlands Antilles
2  ALB      Albania                   NaN
3  ALG      Algeria                   NaN
4  AND      Andorra                   NaN
```

```
# Join the DataFrames

atheletes_df = atheletes.merge(regions, how = 'left', on = 'NOC')
atheletes_df.head()
```

```
   ID                      Name Sex   Age  Height  Weight
Team  \
0   1                 A Dijiang   M  24.0   180.0    80.0
China
1   2                  A Lamusi   M  23.0   170.0    60.0
China
2   3       Gunnar Nielsen Aaby   M  24.0     NaN     NaN
Denmark
3   4      Edgar Lindenau Aabye   M  34.0     NaN     NaN
Denmark/Sweden
4   5  Christine Jacoba Aaftink   F  21.0   185.0    82.0
Netherlands

   NOC        Games  Year  Season       City          Sport  \
0  CHN  1992 Summer  1992  Summer  Barcelona     Basketball
1  CHN  2012 Summer  2012  Summer     London           Judo
2  DEN  1920 Summer  1920  Summer  Antwerpen       Football
3  DEN  1900 Summer  1900  Summer      Paris     Tug-Of-War
4  NED  1988 Winter  1988  Winter    Calgary  Speed Skating

                              Event Medal       region notes
0       Basketball Men's Basketball   NaN        China   NaN
1      Judo Men's Extra-Lightweight   NaN        China   NaN
2           Football Men's Football   NaN      Denmark   NaN
3       Tug-Of-War Men's Tug-Of-War  Gold      Denmark   NaN
4  Speed Skating Women's 500 metres   NaN  Netherlands   NaN
```

```
atheletes_df.shape
```

```
(271116, 17)
```

```
# Column Names Consistent

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

```
   ID                      Name Sex   Age  Height  Weight
Team  \
0   1                 A Dijiang   M  24.0   180.0    80.0
China
```

```
1   2                  A Lamusi   M  23.0   170.0    60.0
China
2   3       Gunnar Nielsen Aaby   M  24.0     NaN     NaN
Denmark
3   4      Edgar Lindenau Aabye   M  34.0     NaN     NaN
Denmark/Sweden
4   5  Christine Jacoba Aaftink   F  21.0   185.0    82.0
Netherlands

   NOC        Games  Year  Season       City          Sport  \
0  CHN  1992 Summer  1992  Summer  Barcelona     Basketball
1  CHN  2012 Summer  2012  Summer     London           Judo
2  DEN  1920 Summer  1920  Summer  Antwerpen       Football
3  DEN  1900 Summer  1900  Summer      Paris     Tug-Of-War
4  NED  1988 Winter  1988  Winter    Calgary  Speed Skating

                              Event Medal       Region Notes
0       Basketball Men's Basketball   NaN        China   NaN
1      Judo Men's Extra-Lightweight   NaN        China   NaN
2           Football Men's Football   NaN      Denmark   NaN
3       Tug-Of-War Men's Tug-Of-War  Gold      Denmark   NaN
4  Speed Skating Women's 500 metres   NaN  Netherlands   NaN
```

```
atheletes_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
```

```
atheletes_df.describe()
```

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

                Year
count  271116.000000
mean     1978.378480
std        29.877632
min      1896.000000
25%      1960.000000
50%      1988.000000
75%      2002.000000
max      2016.000000
```

```
# Checking Null Values

nan_values = atheletes_df.isna()
nan_columns = nan_values.any()
nan_columns
```

```
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
```

```
atheletes_df.isnull().sum()
```

```
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
```

```
# Print the column names containing null values in list format

atheletes_df.columns[atheletes_df.isnull().any()].tolist()
```

```
['Age', 'Height', 'Weight', 'Medal', 'Region', 'Notes']
```

```
# India details

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

```
      ID                             Name Sex   Age  Height  Weight
Team  \
505  281                   S. Abdul Hamid   M   NaN     NaN     NaN
India
506  281                   S. Abdul Hamid   M   NaN     NaN     NaN
India
895  512  Shiny Kurisingal Abraham-Wilson   F  19.0   167.0    53.0
India
896  512  Shiny Kurisingal Abraham-Wilson   F  19.0   167.0    53.0
India
897  512  Shiny Kurisingal Abraham-Wilson   F  23.0   167.0    53.0
India

     NOC        Games  Year  Season         City      Sport  \
505  IND  1928 Summer  1928  Summer    Amsterdam  Athletics
506  IND  1928 Summer  1928  Summer    Amsterdam  Athletics
895  IND  1984 Summer  1984  Summer  Los Angeles  Athletics
896  IND  1984 Summer  1984  Summer  Los Angeles  Athletics
897  IND  1988 Summer  1988  Summer        Seoul  Athletics

                                      Event Medal Region Notes
505      Athletics Men's 110 metres Hurdles   NaN  India   NaN
506      Athletics Men's 400 metres Hurdles   NaN  India   NaN
895            Athletics Women's 800 metres   NaN  India   NaN
896  Athletics Women's 4 x 400 metres Relay   NaN  India   NaN
897            Athletics Women's 800 metres   NaN  India   NaN
```

```
# Japan details

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

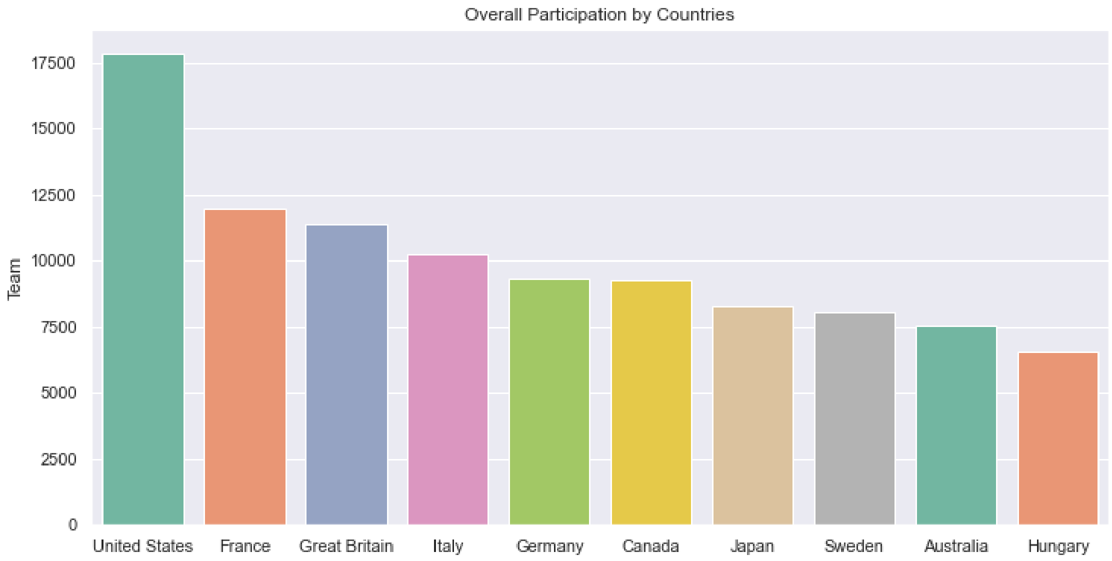
```
      ID          Name Sex   Age  Height  Weight   Team  NOC       
Games  \
625  362  Isao Ko Abe   M  24.0   177.0    75.0  Japan  JPN  1936 
Summer   
629  363   Kazumi Abe   M  28.0   178.0    67.0  Japan  JPN  1976 
Winter   
630  364    Kazuo Abe   M  25.0   166.0    69.0  Japan  JPN  1960 
Summer   
631  365    Kinya Abe   M  23.0   168.0    68.0  Japan  JPN  1992 
Summer   
632  366  Kiyoshi Abe   M  25.0   167.0    62.0  Japan  JPN  1972 
Summer   

     Year  Season       City      Sport  \
625  1936  Summer     Berlin  Athletics   
629  1976  Winter  Innsbruck  Bobsleigh   
630  1960  Summer       Roma  Wrestling   
631  1992  Summer  Barcelona    Fencing   
632  1972  Summer     Munich  Wrestling   

                                        Event Medal Region Notes  
625              Athletics Men's Hammer Throw   NaN  Japan   NaN  
629                     Bobsleigh Men's Four   NaN  Japan   NaN  
630    Wrestling Men's Lightweight, Freestyle   NaN  Japan   NaN  
631              Fencing Men's Foil, Individual   NaN  Japan   NaN  
632  Wrestling Men's Featherweight, Freestyle   NaN  Japan   NaN  
```

```
# Top Countries Participating

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

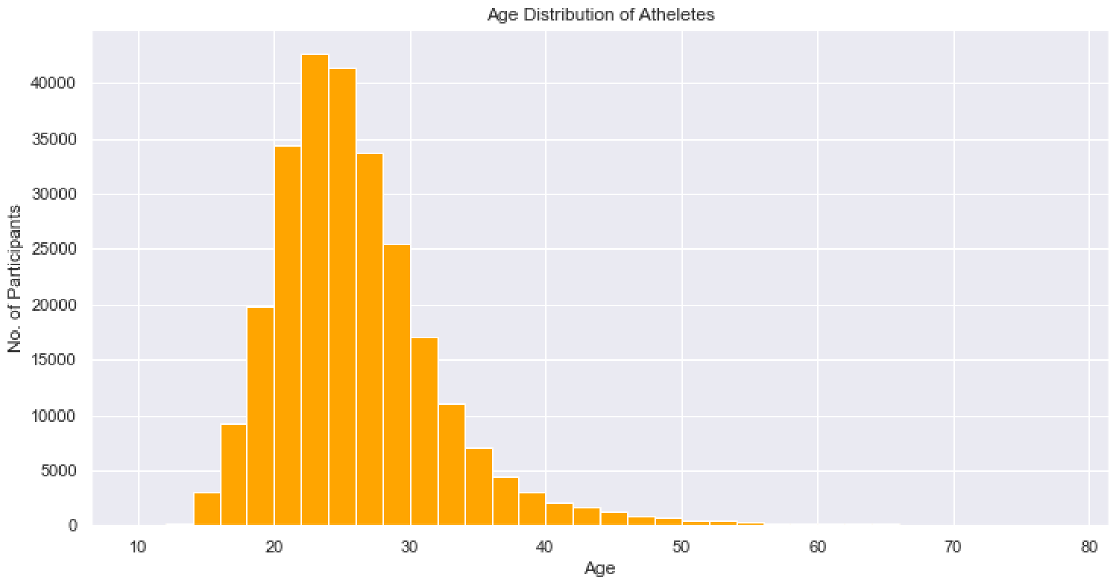
```
United States    17847
France           11988
Great Britain    11404
Italy            10260
Germany           9326
Canada            9279
Japan             8289
Sweden            8052
Australia         7513
Hungary           6547
Name: Team, dtype: int64
```

```
# Plot for Top 10 Countries
```

```
plt.figure(figsize=(12,6))
#plt.xsticks(rotation=20)
plt.title('Overall Participation by Countries')
sns.barplot(x=top_10_countries.index, y=top_10_countries, palette
='Set2');
```



```
# Age Distribution of the Atheletes

plt.figure(figsize=(12,6))
plt.title('Age Distribution of Atheletes')
plt.xlabel('Age')
plt.ylabel('No. of Participants')
plt.hist(atheletes_df.Age, bins = np.arange(10,80,2), color =
'orange', edgecolor = 'white');
```



```
# Summer Olympic Games

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

```
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)
```

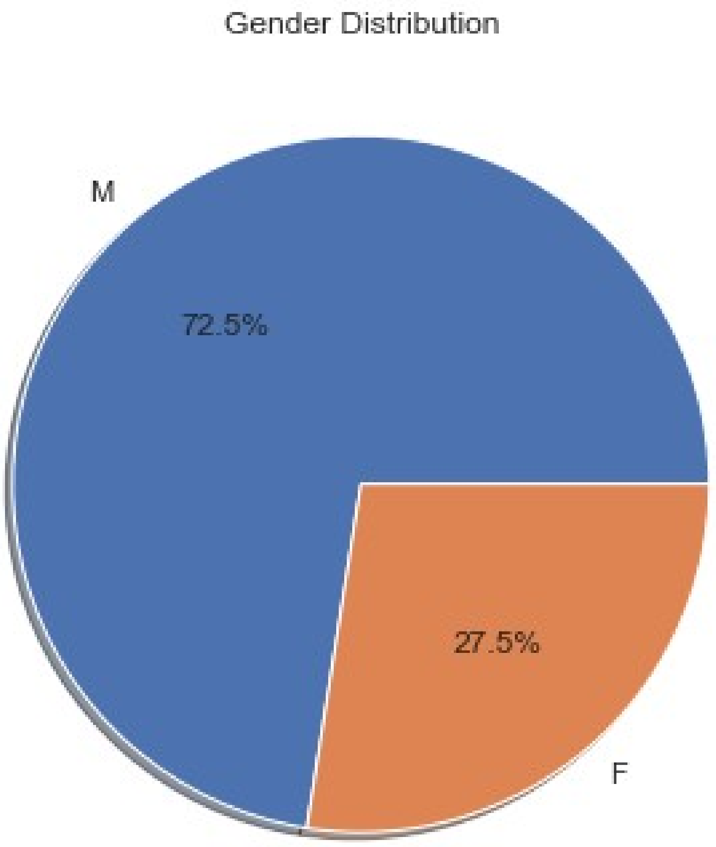
```
# Winter Olympic Games

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

```
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)
```

```
# Male & Female Participants

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

```
M    196594
F     74522
Name: Sex, dtype: int64
```

```
# Pie plot for Male & Female atheletes

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



```
# Total Medals

atheletes_df.Medal.value_counts()
```

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

```
# Total No. of Female Atheletes in each Olympics

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

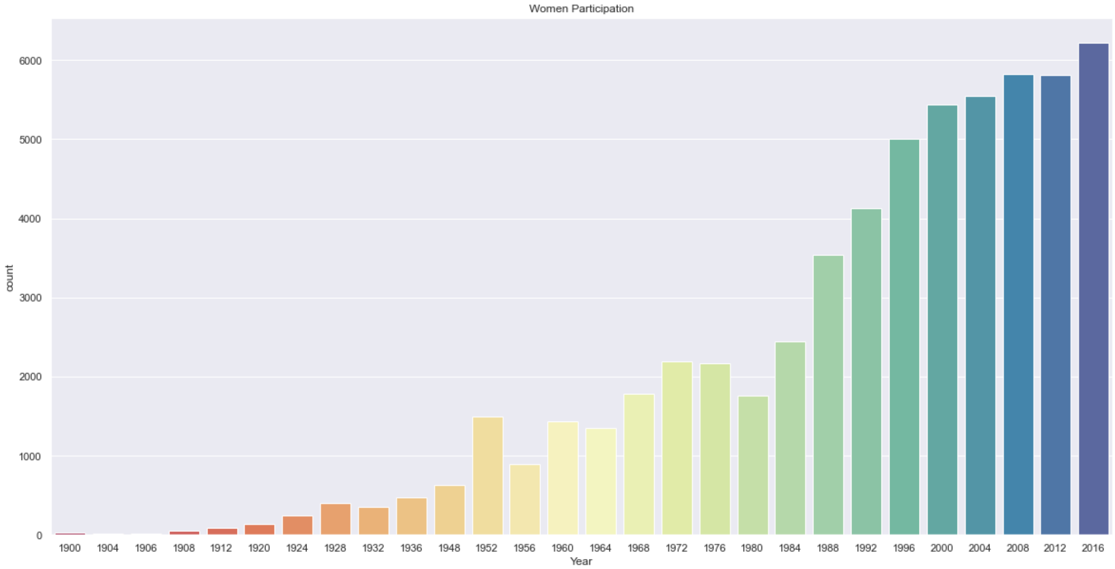
```
    Year   Sex
23  2000  5431
24  2004  5546
25  2008  5816
```

```
26  2012  5815
27  2016  6223
```

```
WomenOlympics = atheletes_df[(atheletes_df.Sex == 'F') &
(atheletes_df.Season == 'Summer')]
```

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

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



```
# Gold Medal Atheletes
```

```
goldmedals = atheletes_df[(atheletes_df.Medal == 'Gold')]
goldmedals.head()
```

```
    ID                     Name Sex   Age  Height  Weight
Team  \
3    4     Edgar Lindenau Aabye   M  34.0     NaN     NaN
Denmark/Sweden
42  17  Paavo Johannes Aaltonen   M  28.0   175.0    64.0
Finland
44  17  Paavo Johannes Aaltonen   M  28.0   175.0    64.0
Finland
48  17  Paavo Johannes Aaltonen   M  28.0   175.0    64.0
Finland
60  20       Kjetil Andr Aamodt   M  20.0   176.0    85.0
Norway

    NOC        Games  Year  Season         City         Sport  \
3   DEN  1900 Summer  1900  Summer        Paris     Tug-Of-War
```

```
42  FIN  1948 Summer  1948  Summer       London     Gymnastics
44  FIN  1948 Summer  1948  Summer       London     Gymnastics
48  FIN  1948 Summer  1948  Summer       London     Gymnastics
60  NOR  1992 Winter  1992  Winter  Albertville  Alpine Skiing

                               Event Medal   Region Notes
3        Tug-Of-War Men's Tug-Of-War  Gold  Denmark   NaN
42  Gymnastics Men's Team All-Around  Gold  Finland   NaN
44      Gymnastics Men's Horse Vault  Gold  Finland   NaN
48  Gymnastics Men's Pommelled Horse  Gold  Finland   NaN
60       Alpine Skiing Men's Super G  Gold   Norway   NaN
```

```
# Gold Medals earned by 60+
```

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

```
6
```

```
sport_event = goldmedals['Sport'][goldmedals['Age'] > 60]
sport_event
```

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

```
# Gold Medals for each Country
```

```
goldmedals.Region.value_counts().reset_index(name='Medal').head(10)
```

```
     index  Medal
0      USA   2638
1   Russia   1599
2  Germany   1301
3       UK    678
4    Italy    575
5   France    501
6   Sweden    479
7   Canada    463
8  Hungary    432
9   Norway    378
```

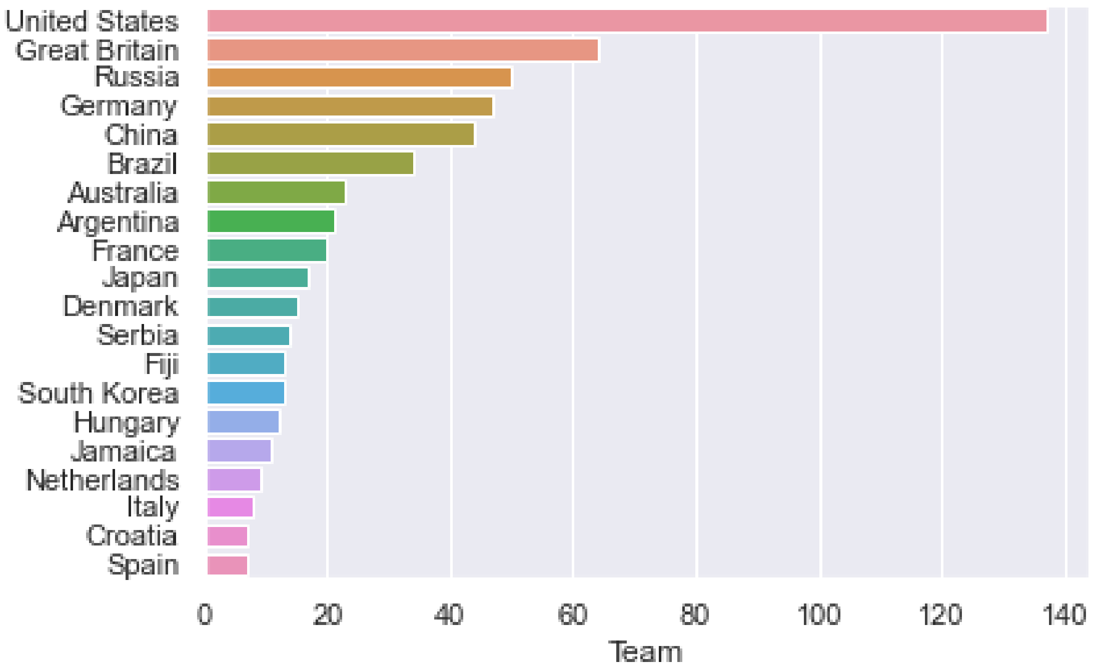
```
# Rio Olympics
```

```
max_year = atheletes_df.Year.max()
print(max_year)

team_names = atheletes_df[(atheletes_df.Year == max_year) &
(atheletes_df.Medal == 'Gold')].Team
team_names.value_counts().head(20)
```

2016

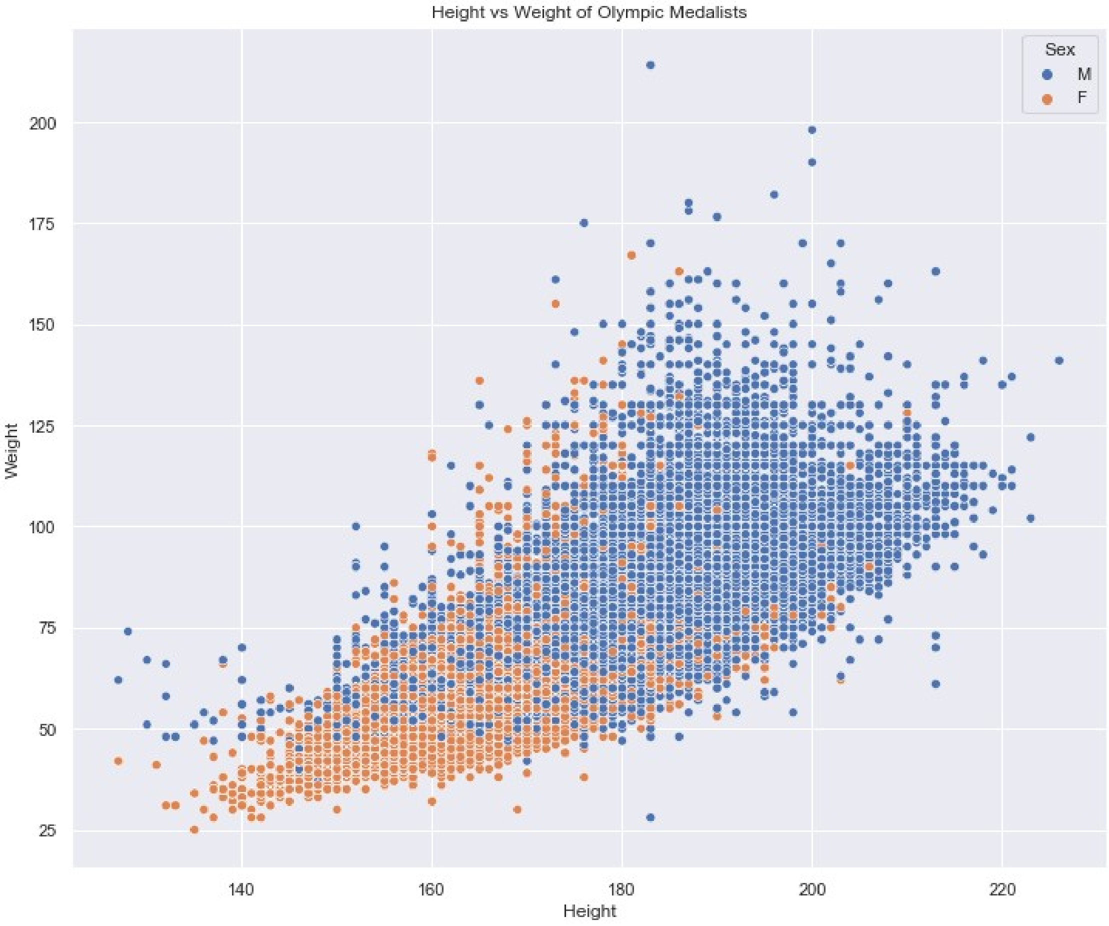
```
United States    137
Great Britain     64
Russia            50
Germany           47
China             44
Brazil            34
Australia         23
Argentina         21
France            20
Japan             17
Denmark           15
Serbia            14
Fiji              13
South Korea       13
Hungary           12
Jamaica           11
Netherlands        9
Italy              8
Croatia            7
Spain              7
Name: Team, dtype: int64
```

```
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 Rio Olympics");
```



```
# Extracting Data without Null values

not_null_medals = atheletes_df[(atheletes_df['Height'].notnull()) &
(atheletes_df['Weight'].notnull())]

# Scatter Plot

plt.figure(figsize=(12,10))
axis = sns.scatterplot(x="Height", y="Weight", data=not_null_medals,
hue = 'Sex')
plt.title('Height vs Weight of Olympic Medalists')
```

```
Text(0.5, 1.0, 'Height vs Weight of Olympic Medalists')
```


Height vs Weight of Olympic Medalists
Sex
M
F
Weight
200
175
150
125
100
75
50
25
Height
140
160
180
200
220