In [1]:

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Load the dataset

In [6]:

```
df = pd.read_csv('athlete_events.csv')
```

Data Cleaning

Handle missing values

In [7]:

```
missing_values = df.isnull().sum()
```

In [8]:

```
duplicate_rows = df.duplicated().sum()
```

Exploratory Data Analysis (EDA)

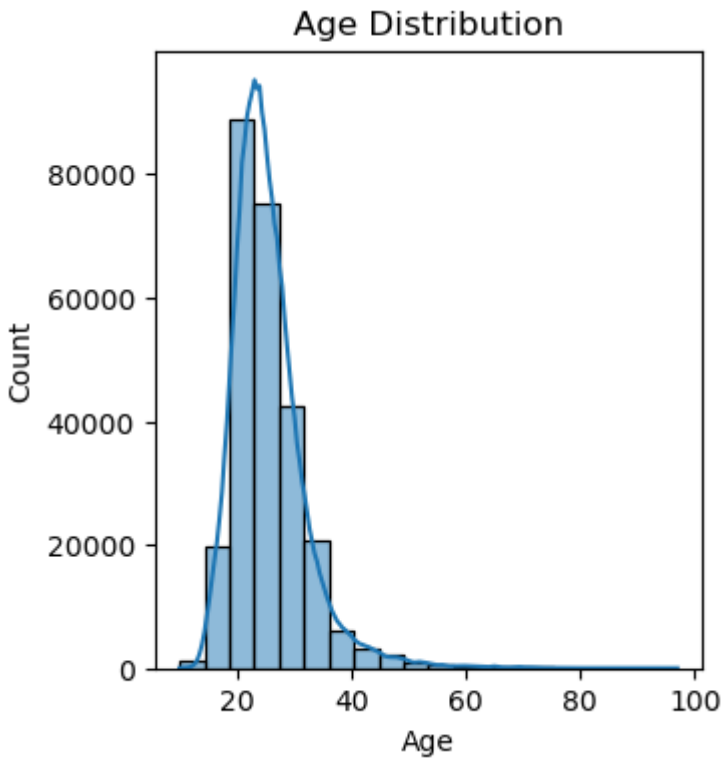
Age Distribution

In [9]:

```
plt.figure(figsize=(12, 4))
plt.subplot(1, 3, 1)
sns.histplot(df['Age'].dropna(), bins=20, kde=True)
plt.title('Age Distribution')
```

Out[9]:

```
Text(0.5, 1.0, 'Age Distribution')
```



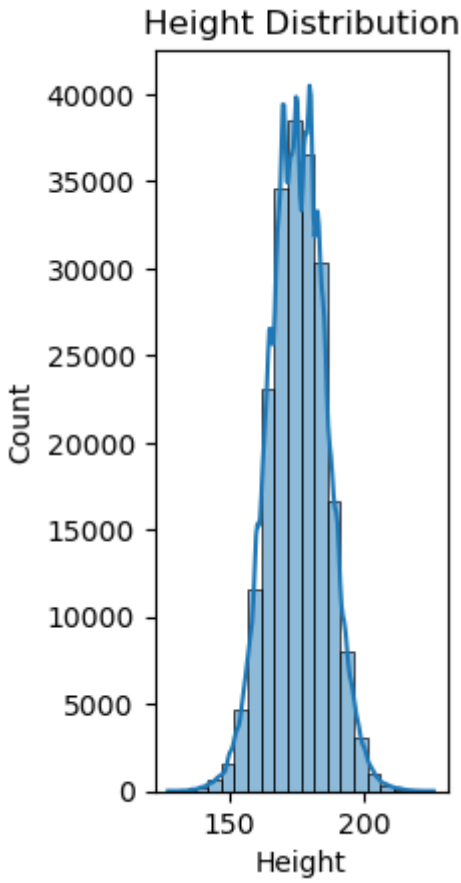
Height Distribution

In [10]:

```
plt.subplot(1, 3, 2)
sns.histplot(df['Height'].dropna(), bins=20, kde=True)
plt.title('Height Distribution')
```

Out[10]:

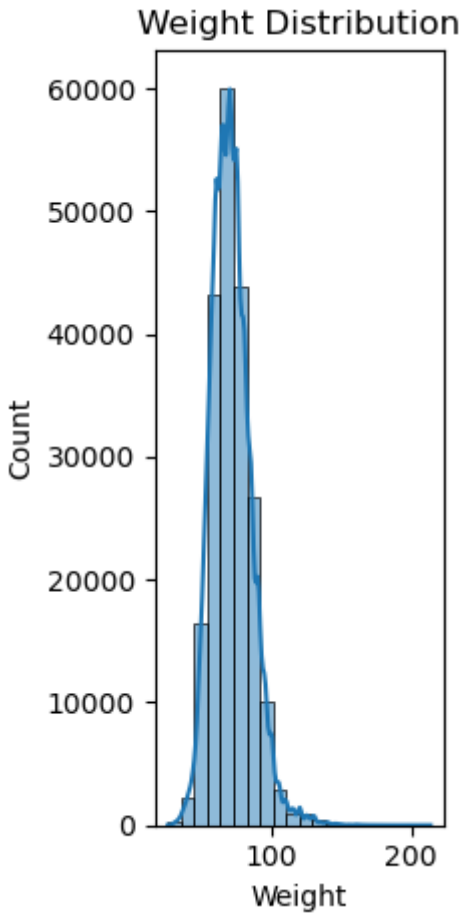
```
Text(0.5, 1.0, 'Height Distribution')
```



Weight Distribution

In [12]:

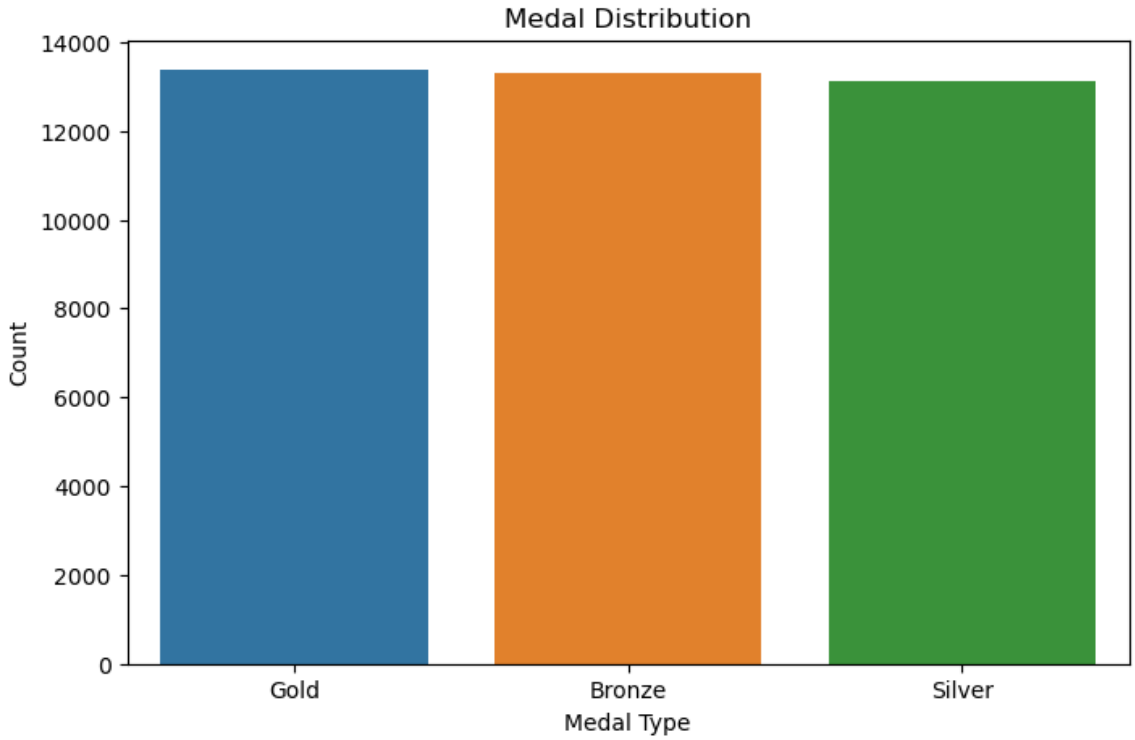
```
plt.subplot(1, 3, 3)
sns.histplot(df['Weight'].dropna(), bins=20, kde=True)
plt.title('Weight Distribution')

plt.tight_layout()
plt.show()
```



Medal Distribution

In [13]:

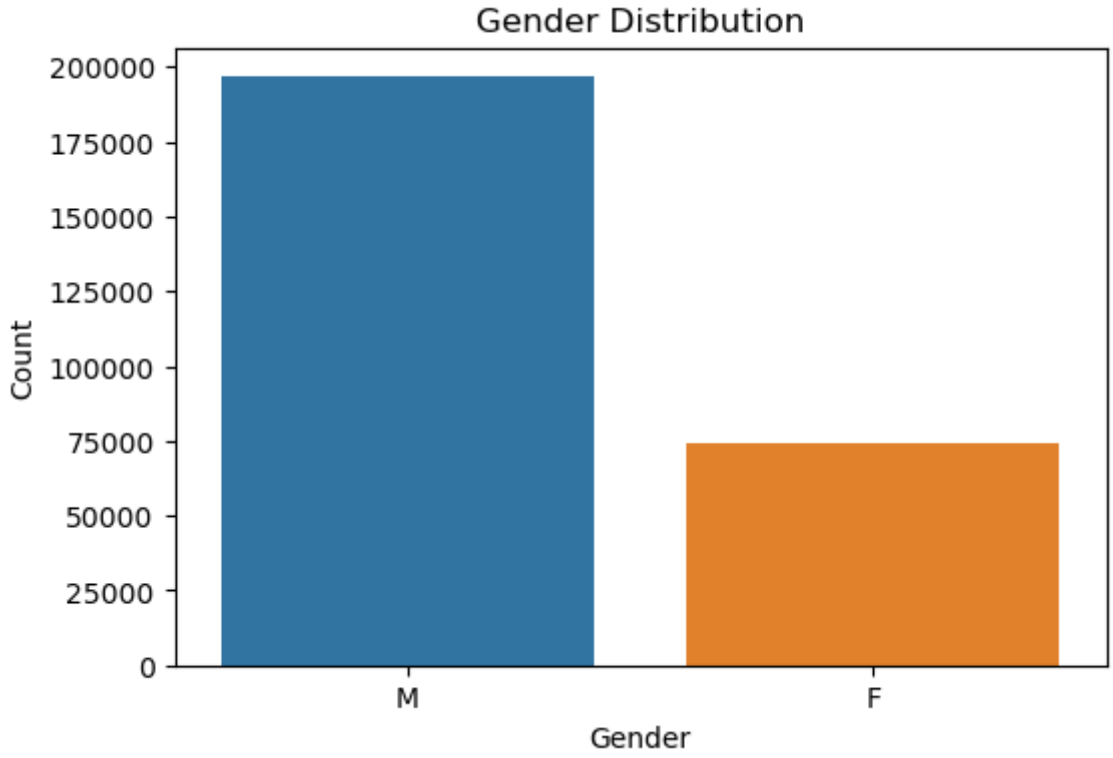
```
medal_counts = df['Medal'].value_counts()
plt.figure(figsize=(8, 5))
sns.barplot(x=medal_counts.index, y=medal_counts.values)
plt.xlabel('Medal Type')
plt.ylabel('Count')
plt.title('Medal Distribution')
plt.show()
```



Gender Distribution

In [14]:

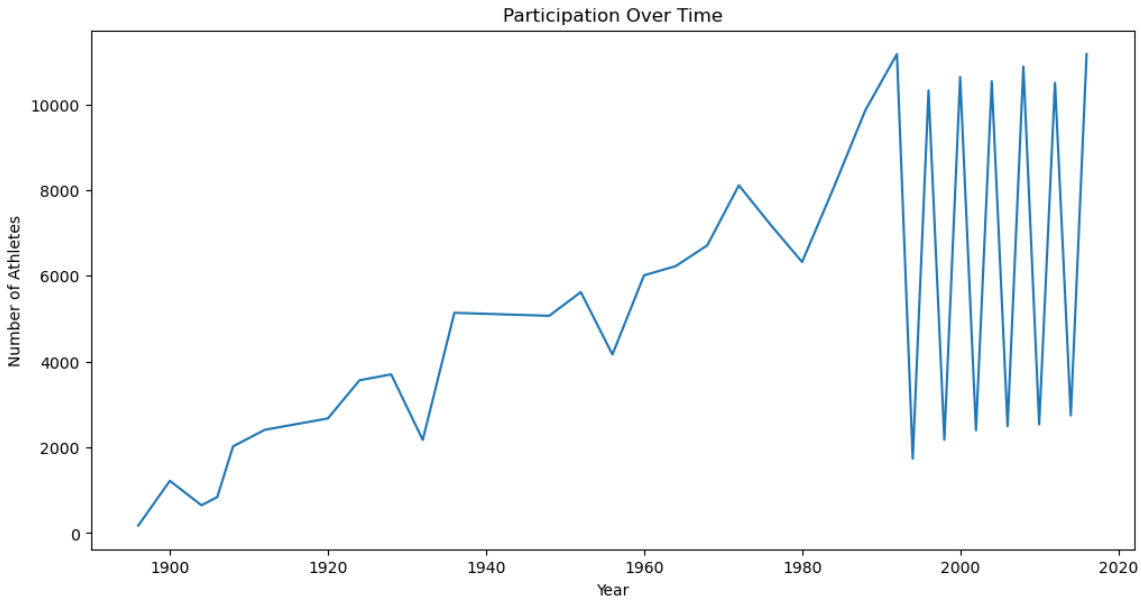
```
gender_counts = df['Sex'].value_counts()
plt.figure(figsize=(6, 4))
sns.barplot(x=gender_counts.index, y=gender_counts.values)
plt.xlabel('Gender')
plt.ylabel('Count')
plt.title('Gender Distribution')
plt.show()
```



Participation Over Time

In [15]:

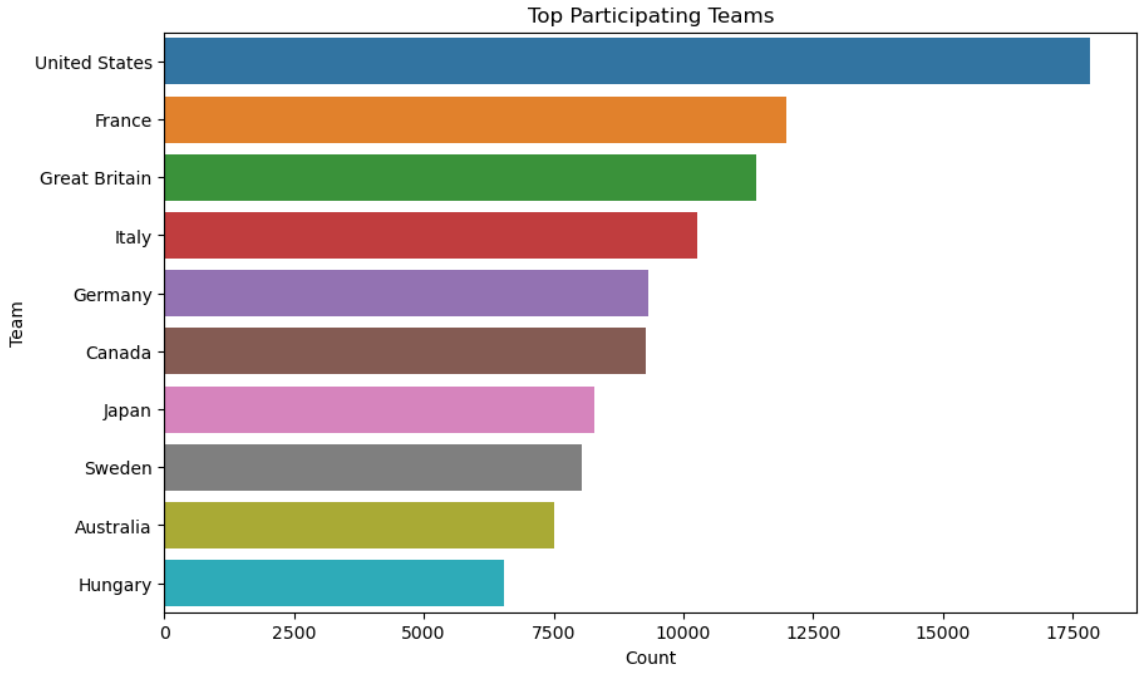
```
yearly_participation = df.groupby('Year')['Name'].nunique()
plt.figure(figsize=(12, 6))
sns.lineplot(x=yearly_participation.index, y=yearly_participation.values)
plt.xlabel('Year')
plt.ylabel('Number of Athletes')
plt.title('Participation Over Time')
plt.show()
```



Top Participating Teams

In [17]:

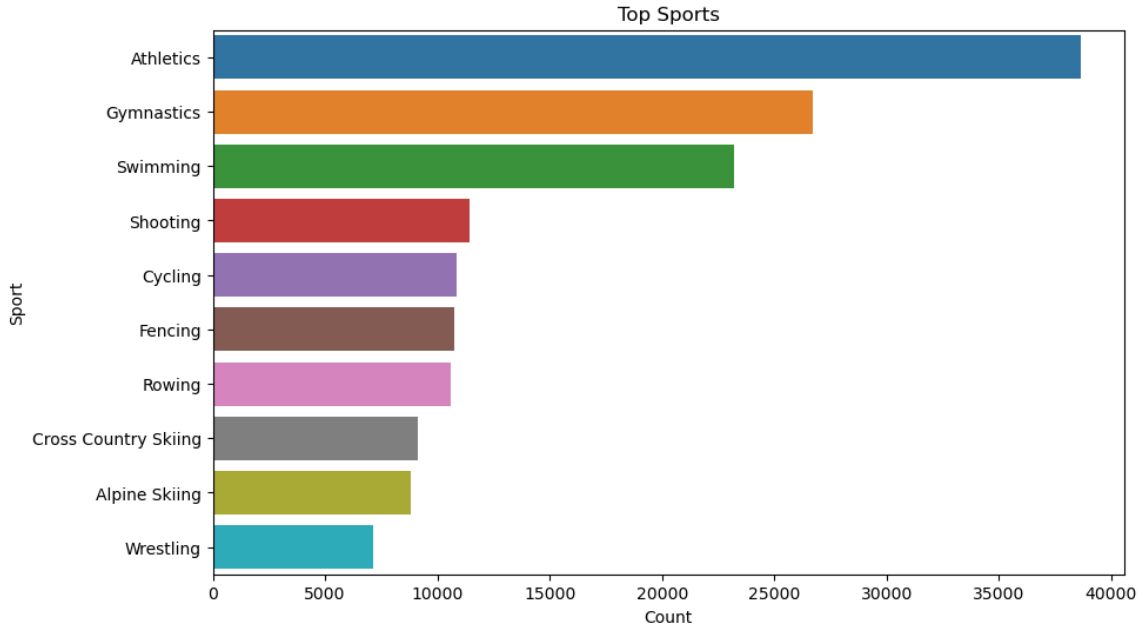
```
top_teams = df['Team'].value_counts().head(10)
plt.figure(figsize=(10, 6))
sns.barplot(x=top_teams.values, y=top_teams.index)
plt.xlabel('Count')
plt.ylabel('Team')
plt.title('Top Participating Teams')
plt.show()
```



Top Sports

In [18]:

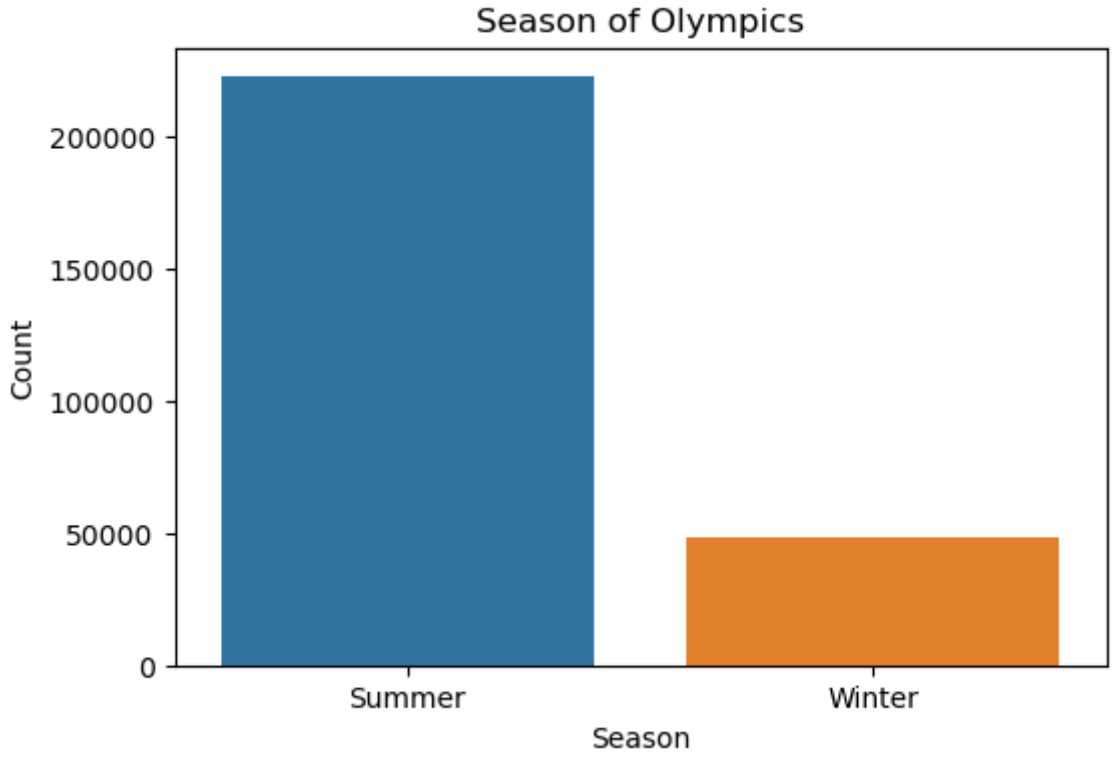
```
 top_sports = df['Sport'].value_counts().head(10)
plt.figure(figsize=(10, 6))
sns.barplot(x=top_sports.values, y=top_sports.index)
plt.xlabel('Count')
plt.ylabel('Sport')
plt.title('Top Sports')
plt.show()
```



Season of Olympics

In [19]:

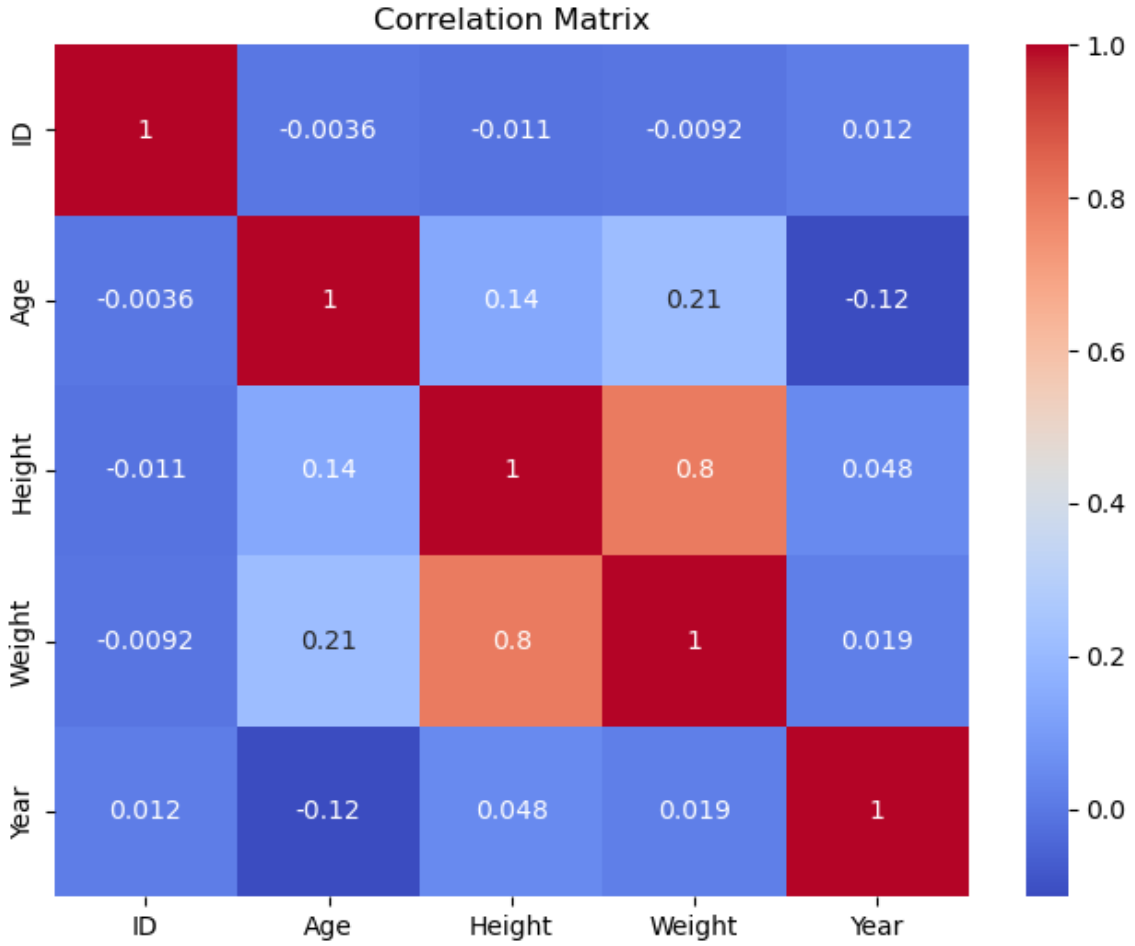
```
season_counts = df['Season'].value_counts()
plt.figure(figsize=(6, 4))
sns.barplot(x=season_counts.index, y=season_counts.values)
plt.xlabel('Season')
plt.ylabel('Count')
plt.title('Season of Olympics')
plt.show()
```



Correlation Matrix

In [20]:

```
correlation_matrix = df.corr()
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```



In [ ]: