Project Overview: Olympic History Dataset Analysis

The Olympic History dataset, covering the period from Athens 1896 to Rio 2016, serves as a comprehensive repository of the modern Olympic Games' evolution. It encapsulates data on athlete participation, country performance, and event details, offering a rich resource for exploring historical trends and patterns.

Objectives:

1. Explore Participation Trends:

Investigate changes in athlete participation over time, including the emergence of new sporting disciplines and shifts in gender representation.

2. Analyze Performance Dynamics:

Examine the performance of nations across different Olympics editions, identifying powerhouse countries and evaluating their dominance in specific sports.

3. Gender Representation:

Assess the progress of gender equality in Olympic sports by analyzing participation rates and medal distributions among male and female athletes.

4. Sporting Diversity:

Delve into the diversity of Olympic sports and events, uncovering trends in popularity, emergence, and decline of different disciplines over the years.

Data Preparation:

The dataset underwent meticulous preprocessing using Numpy and Pandas libraries to handle missing values, ensure data integrity, and standardize formats. The cleaning process resulted in a refined dataset comprising 39,772 rows and 15 columns, ready for in-depth analysis.

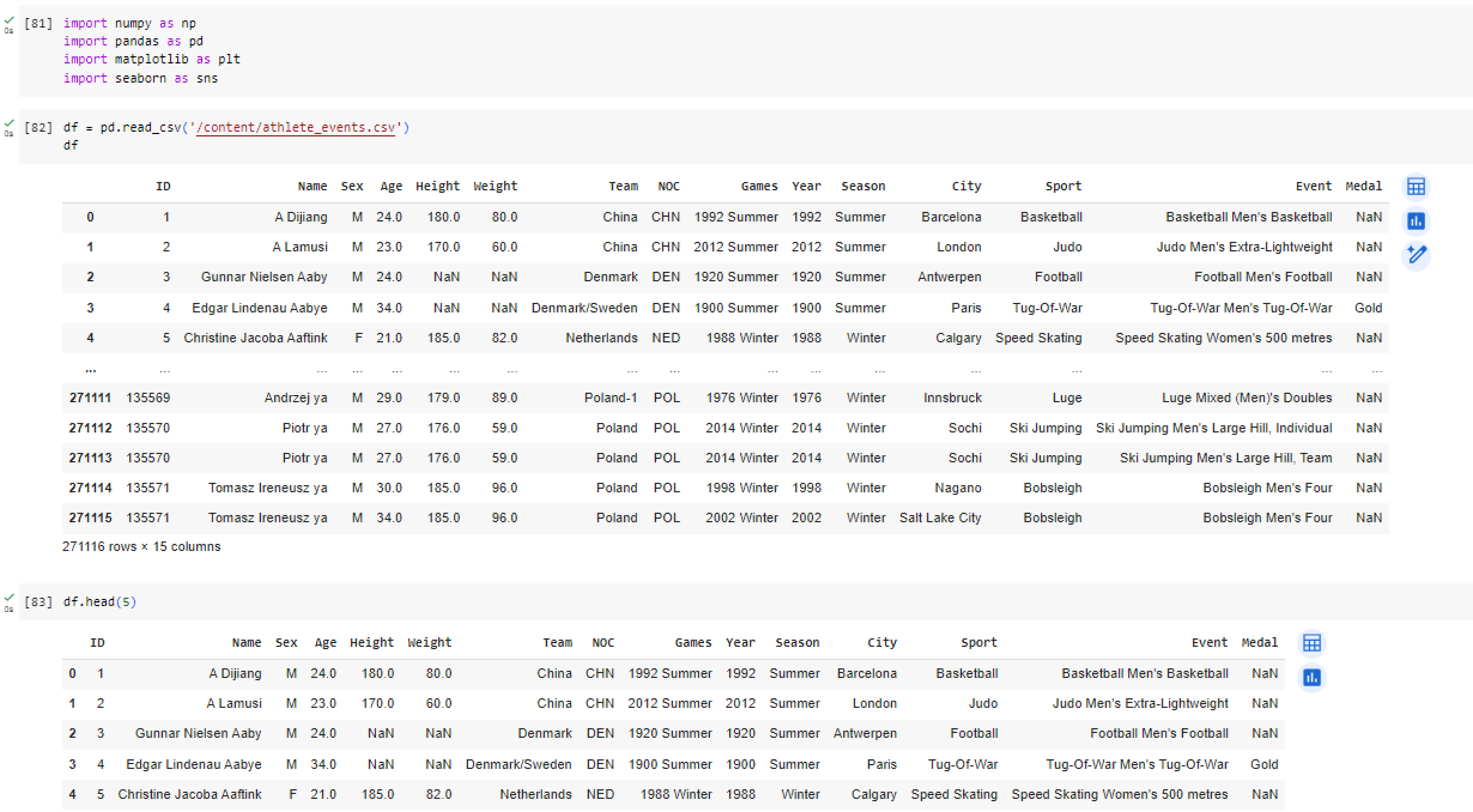
Approach:

1. Exploratory Analysis:

Conducted comprehensive exploratory data analysis (EDA) to unveil underlying patterns, outliers, and correlations within the dataset. Utilized statistical measures and visualization techniques to gain insights.

2. Visualization Techniques:

Leveraged the powerful visualization capabilities of Matplotlib and Seaborn libraries to create informative plots, including bar charts, line plots, pie chart , and scatter plots, to illustrate trends and relationships effectively.



```
[81] import numpy as np
     import pandas as pd
     import matplotlib as plt
     import seaborn as sns
```

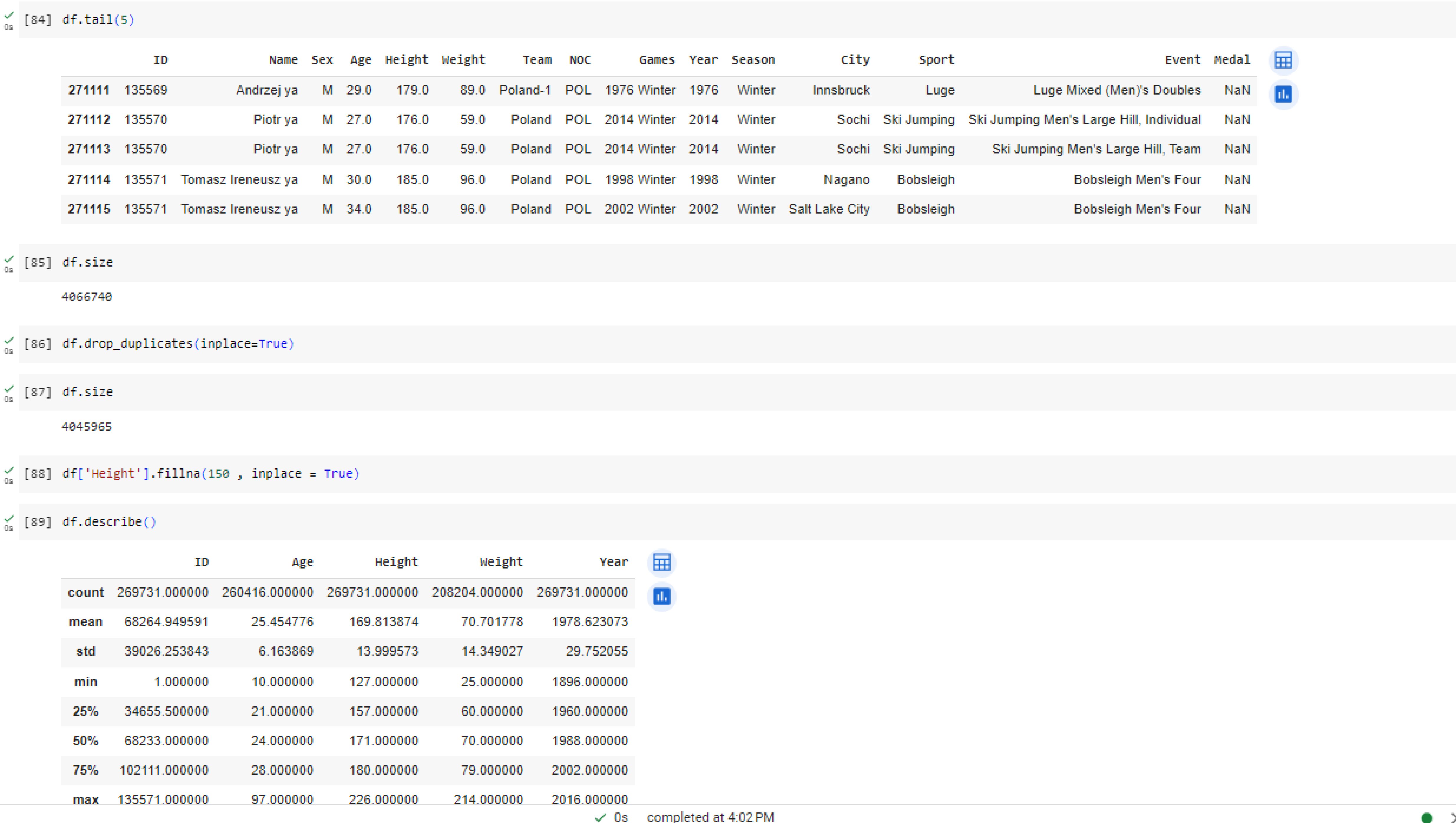
```
[82] df = pd.read_csv('/content/athlete_events.csv')
     df
```

	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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
271111	135569	Andrzej ya	M	29.0	179.0	89.0	Poland-1	POL	1976 Winter	1976	Winter	Innsbruck	Luge	Luge Mixed (Men)'s Doubles	NaN
271112	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014	Winter	Sochi	Ski Jumping	Ski Jumping Men's Large Hill, Individual	NaN
271113	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014	Winter	Sochi	Ski Jumping	Ski Jumping Men's Large Hill, Team	NaN
271114	135571	Tomasz Ireneusz ya	M	30.0	185.0	96.0	Poland	POL	1998 Winter	1998	Winter	Nagano	Bobsleigh	Bobsleigh Men's Four	NaN
271115	135571	Tomasz Ireneusz ya	M	34.0	185.0	96.0	Poland	POL	2002 Winter	2002	Winter	Salt Lake City	Bobsleigh	Bobsleigh Men's Four	NaN

271116 rows × 15 columns

```
[83] df.head(5)
```

	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



```
[84] df.tail(5)
```

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
271111	135569	Andrzej ya	M	29.0	179.0	89.0	Poland-1	POL	1976 Winter	1976	Winter	Innsbruck	Luge	Luge Mixed (Men)'s Doubles	NaN
271112	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014	Winter	Sochi	Ski Jumping	Ski Jumping Men's Large Hill, Individual	NaN
271113	135570	Piotr ya	M	27.0	176.0	59.0	Poland	POL	2014 Winter	2014	Winter	Sochi	Ski Jumping	Ski Jumping Men's Large Hill, Team	NaN
271114	135571	Tomasz Ireneusz ya	M	30.0	185.0	96.0	Poland	POL	1998 Winter	1998	Winter	Nagano	Bobsleigh	Bobsleigh Men's Four	NaN
271115	135571	Tomasz Ireneusz ya	M	34.0	185.0	96.0	Poland	POL	2002 Winter	2002	Winter	Salt Lake City	Bobsleigh	Bobsleigh Men's Four	NaN

```
[85] df.size
```

```
4066740
```

```
[86] df.drop_duplicates(inplace=True)
```

```
[87] df.size
```

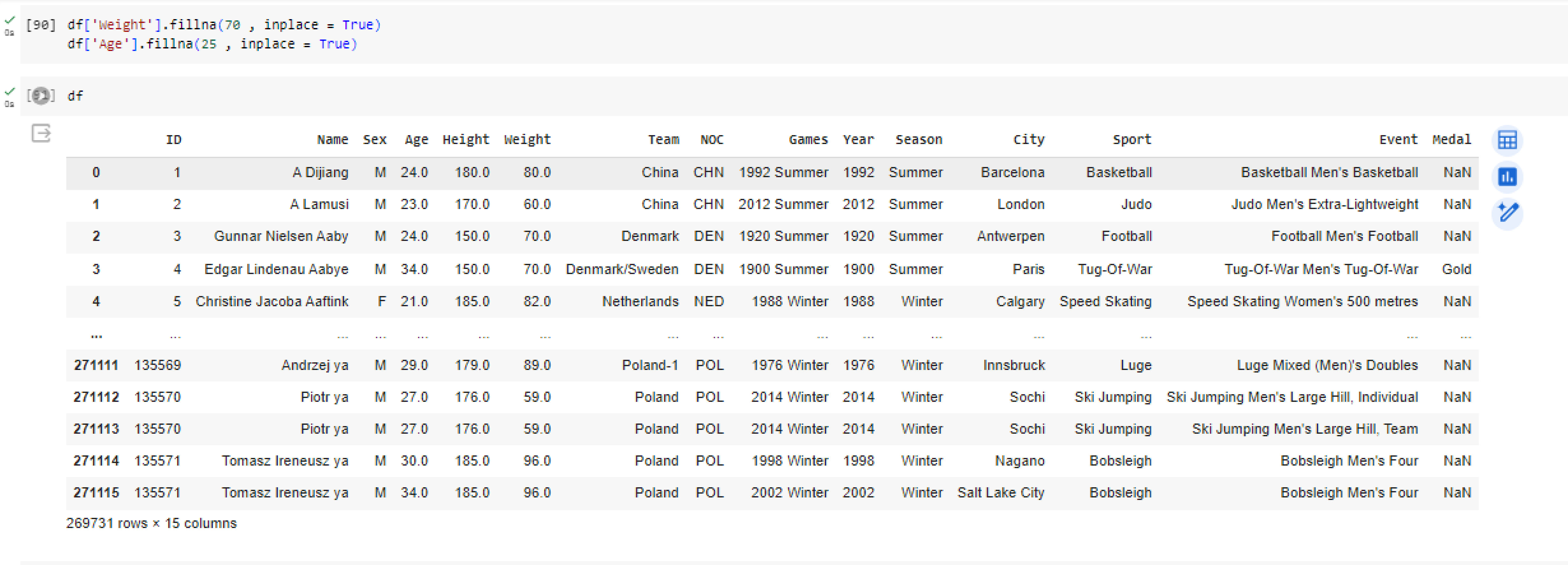
```
4045965
```

```
[88] df['Height'].fillna(150 , inplace = True)
```

```
[89] df.describe()
```

	ID	Age	Height	Weight	Year
count	269731.000000	260416.000000	269731.000000	208204.000000	269731.000000
mean	68264.949591	25.454776	169.813874	70.701778	1978.623073
std	39026.253843	6.163869	13.999573	14.349027	29.752055
min	1.000000	10.000000	127.000000	25.000000	1896.000000
25%	34655.500000	21.000000	157.000000	60.000000	1960.000000
50%	68233.000000	24.000000	171.000000	70.000000	1988.000000
75%	102111.000000	28.000000	180.000000	79.000000	2002.000000
max	135571.000000	97.000000	226.000000	214.000000	2016.000000

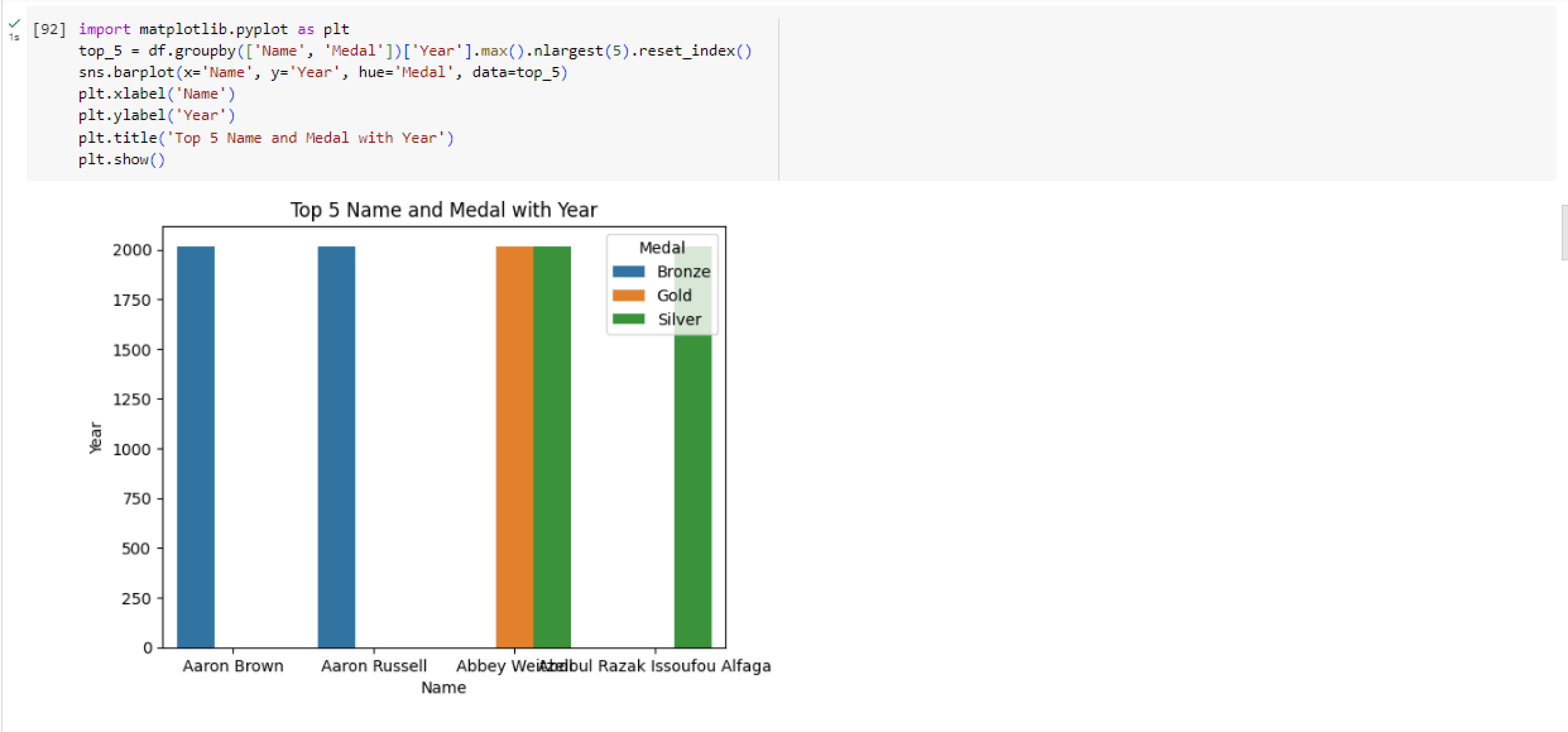
✓ 0s completed at 4:02 PM



```
[90] df['Weight'].fillna(70 , inplace = True)
     df['Age'].fillna(25 , inplace = True)
```

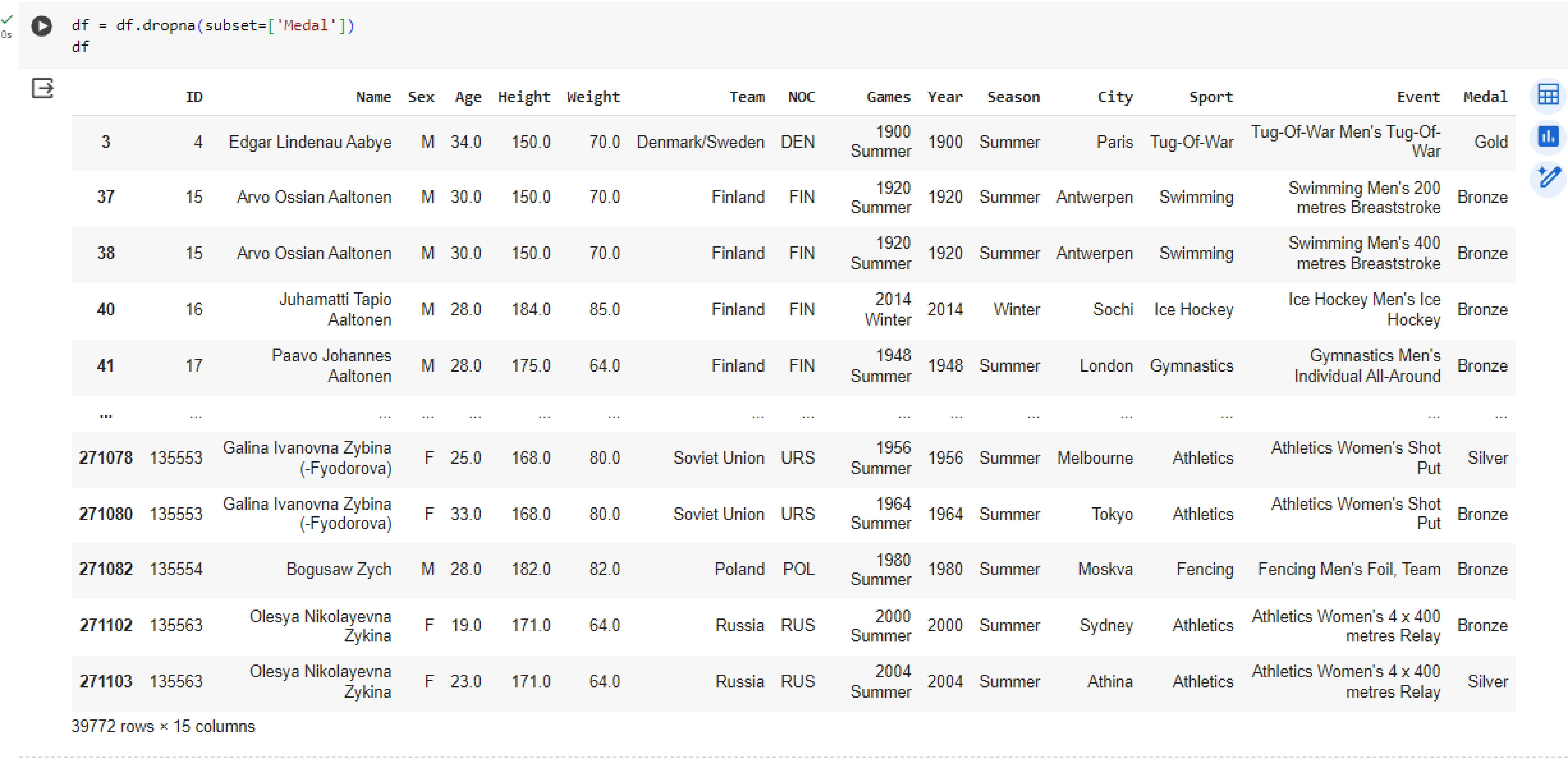
```
[91] df
```

```
             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   150.0    70.0         Denmark  DEN  1920 Summer  1920  Summer       Antwerpen       Football                   Football Men's Football    NaN
     3        4   Edgar Lindenau Aabye   M  34.0   150.0    70.0  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
   ...      ...                    ...  ...   ...     ...     ...             ...  ...          ...   ...     ...             ...            ...                                       ...    ...
271111   135569            Andrzej ya   M  29.0   179.0    89.0        Poland-1  POL  1976 Winter  1976  Winter       Innsbruck           Luge               Luge Mixed (Men)'s Doubles    NaN
271112   135570              Piotr ya   M  27.0   176.0    59.0          Poland  POL  2014 Winter  2014  Winter           Sochi    Ski Jumping  Ski Jumping Men's Large Hill, Individual    NaN
271113   135570              Piotr ya   M  27.0   176.0    59.0          Poland  POL  2014 Winter  2014  Winter           Sochi    Ski Jumping        Ski Jumping Men's Large Hill, Team    NaN
271114   135571     Tomasz Ireneusz ya   M  30.0   185.0    96.0          Poland  POL  1998 Winter  1998  Winter          Nagano      Bobsleigh                      Bobsleigh Men's Four    NaN
271115   135571     Tomasz Ireneusz ya   M  34.0   185.0    96.0          Poland  POL  2002 Winter  2002  Winter  Salt Lake City      Bobsleigh                      Bobsleigh Men's Four    NaN

269731 rows × 15 columns
```



1s [92]

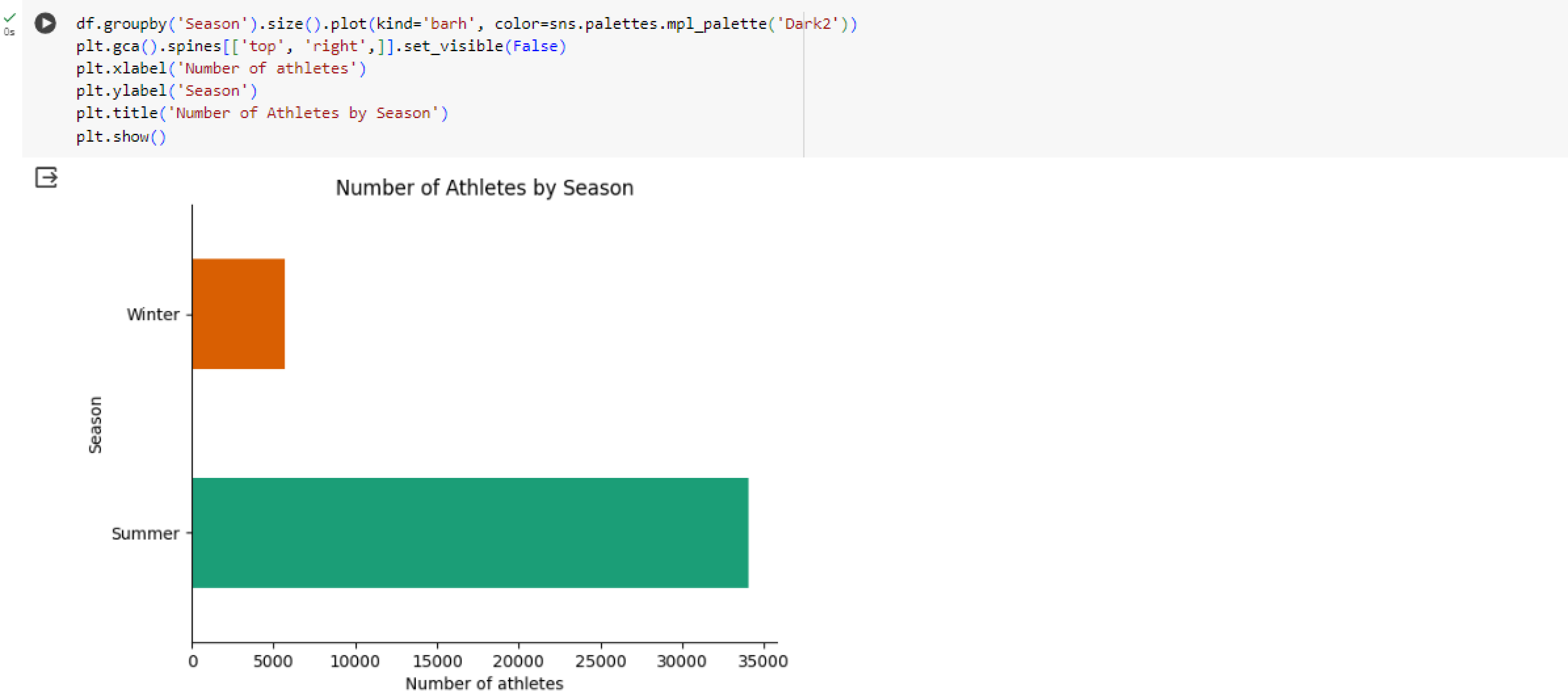
```python
import matplotlib.pyplot as plt
top_5 = df.groupby(['Name', 'Medal'])['Year'].max().nlargest(5).reset_index()
sns.barplot(x='Name', y='Year', hue='Medal', data=top_5)
plt.xlabel('Name')
plt.ylabel('Year')
plt.title('Top 5 Name and Medal with Year')
plt.show()
```



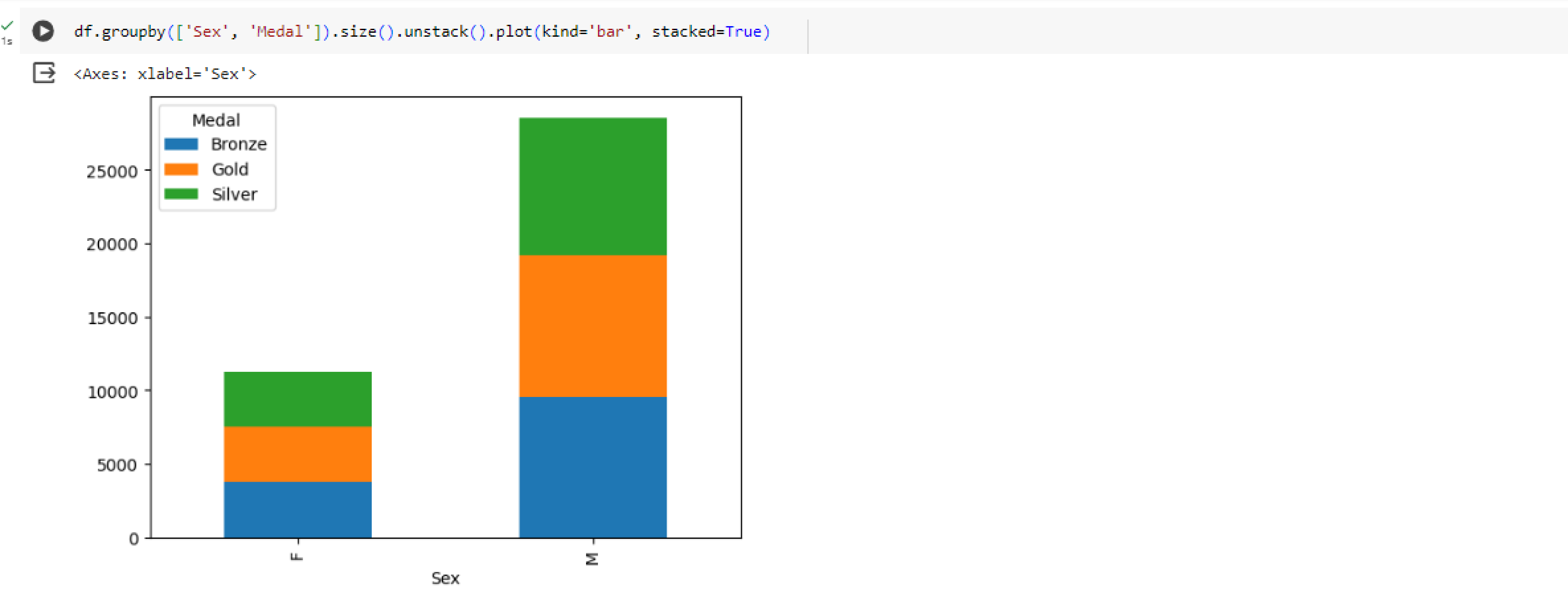
```python
df = df.dropna(subset=['Medal'])
df
```

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
3	4	Edgar Lindenau Aabye	M	34.0	150.0	70.0	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
37	15	Arvo Ossian Aaltonen	M	30.0	150.0	70.0	Finland	FIN	1920 Summer	1920	Summer	Antwerpen	Swimming	Swimming Men's 200 metres Breaststroke	Bronze
38	15	Arvo Ossian Aaltonen	M	30.0	150.0	70.0	Finland	FIN	1920 Summer	1920	Summer	Antwerpen	Swimming	Swimming Men's 400 metres Breaststroke	Bronze
40	16	Juhamatti Tapio Aaltonen	M	28.0	184.0	85.0	Finland	FIN	2014 Winter	2014	Winter	Sochi	Ice Hockey	Ice Hockey Men's Ice Hockey	Bronze
41	17	Paavo Johannes Aaltonen	M	28.0	175.0	64.0	Finland	FIN	1948 Summer	1948	Summer	London	Gymnastics	Gymnastics Men's Individual All-Around	Bronze
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
271078	135553	Galina Ivanovna Zybina (-Fyodorova)	F	25.0	168.0	80.0	Soviet Union	URS	1956 Summer	1956	Summer	Melbourne	Athletics	Athletics Women's Shot Put	Silver
271080	135553	Galina Ivanovna Zybina (-Fyodorova)	F	33.0	168.0	80.0	Soviet Union	URS	1964 Summer	1964	Summer	Tokyo	Athletics	Athletics Women's Shot Put	Bronze
271082	135554	Bogusaw Zych	M	28.0	182.0	82.0	Poland	POL	1980 Summer	1980	Summer	Moskva	Fencing	Fencing Men's Foil, Team	Bronze
271102	135563	Olesya Nikolayevna Zykina	F	19.0	171.0	64.0	Russia	RUS	2000 Summer	2000	Summer	Sydney	Athletics	Athletics Women's 4 x 400 metres Relay	Bronze
271103	135563	Olesya Nikolayevna Zykina	F	23.0	171.0	64.0	Russia	RUS	2004 Summer	2004	Summer	Athina	Athletics	Athletics Women's 4 x 400 metres Relay	Silver

39772 rows × 15 columns

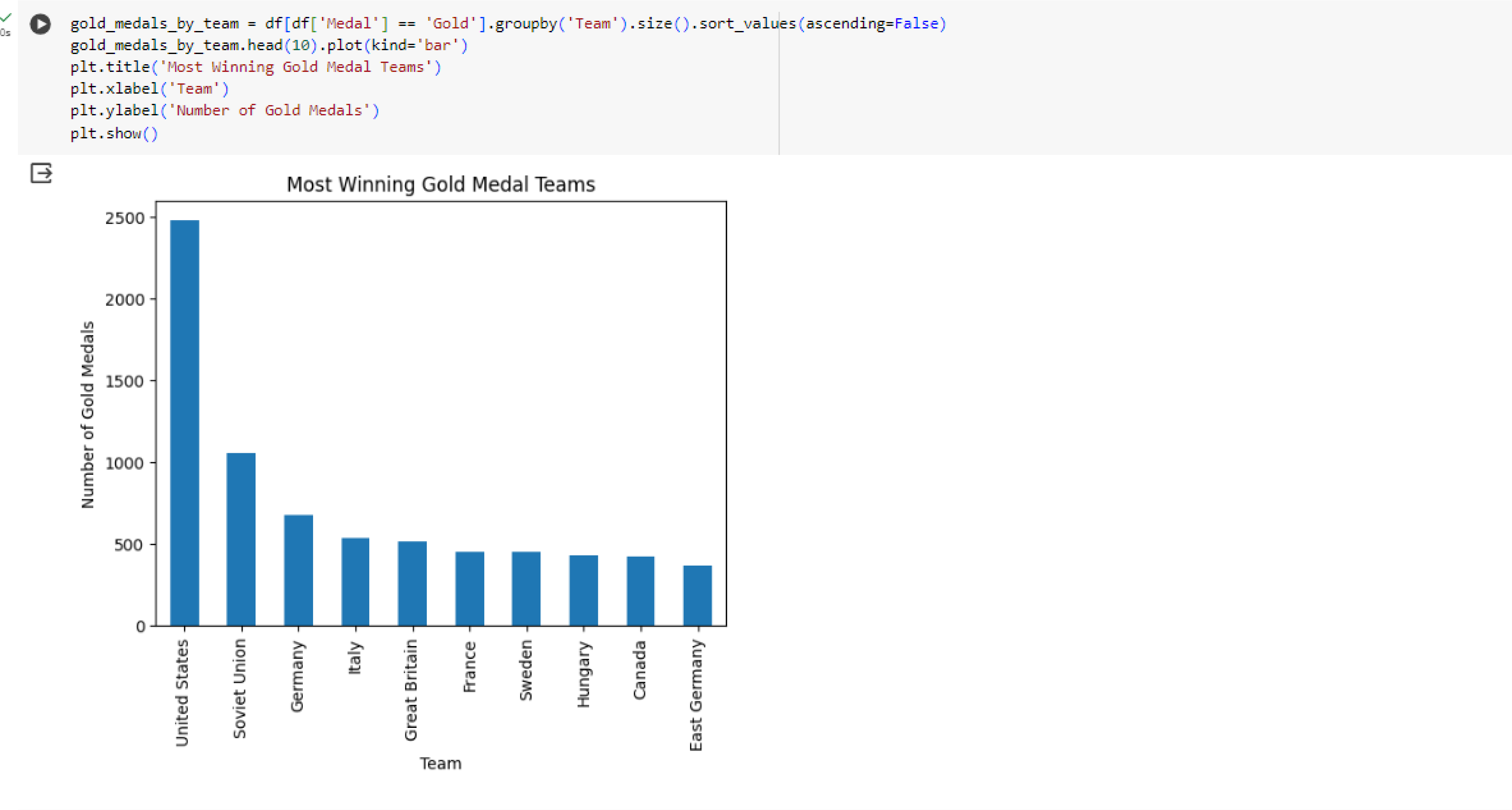


```
df.groupby('Season').size().plot(kind='barh', color=sns.palettes.mpl_palette('Dark2'))
plt.gca().spines[['top', 'right',]].set_visible(False)
plt.xlabel('Number of athletes')
plt.ylabel('Season')
plt.title('Number of Athletes by Season')
plt.show()
```



```
df.groupby(['Sex', 'Medal']).size().unstack().plot(kind='bar', stacked=True)
```

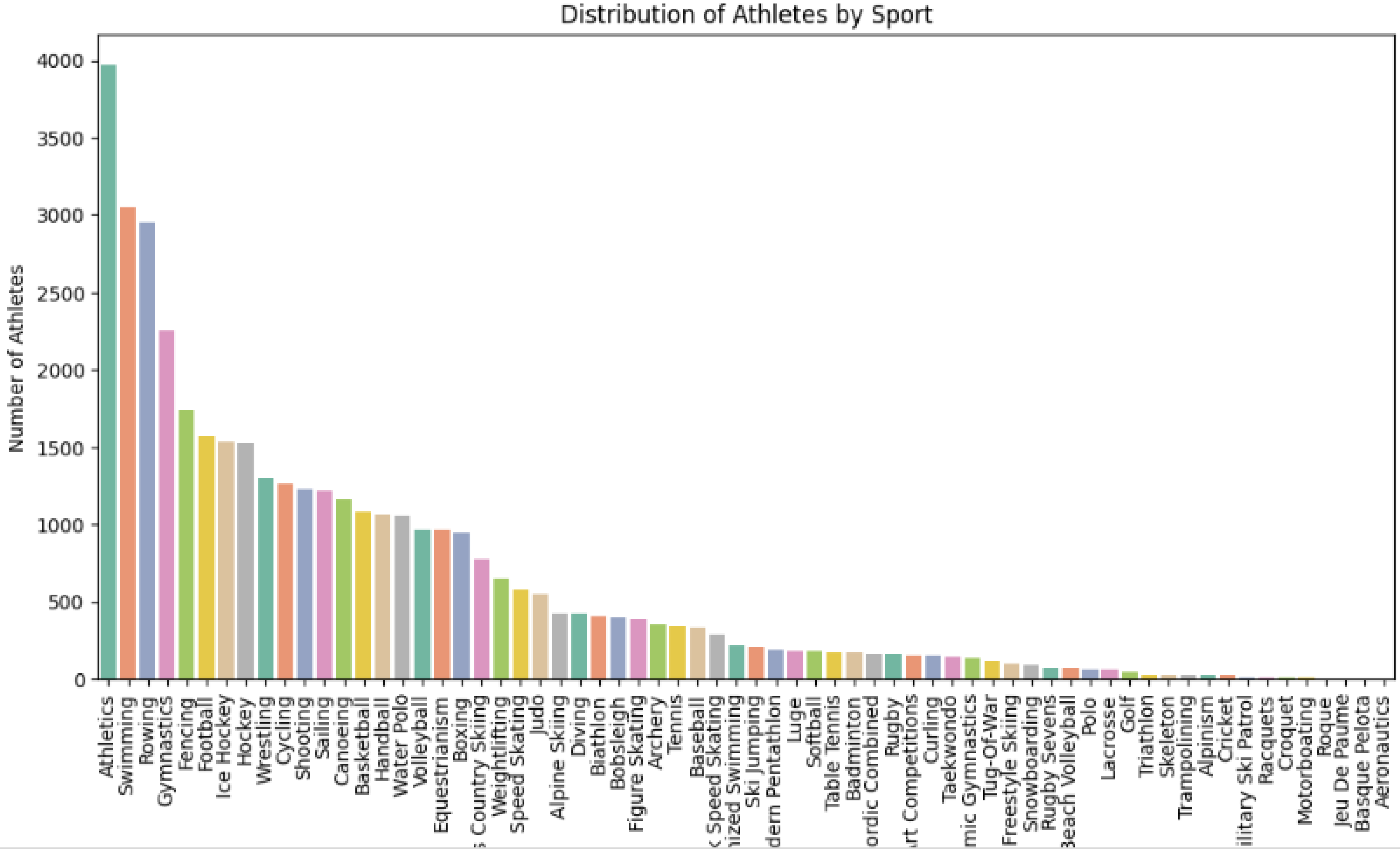
```
<Axes: xlabel='Sex'>
```

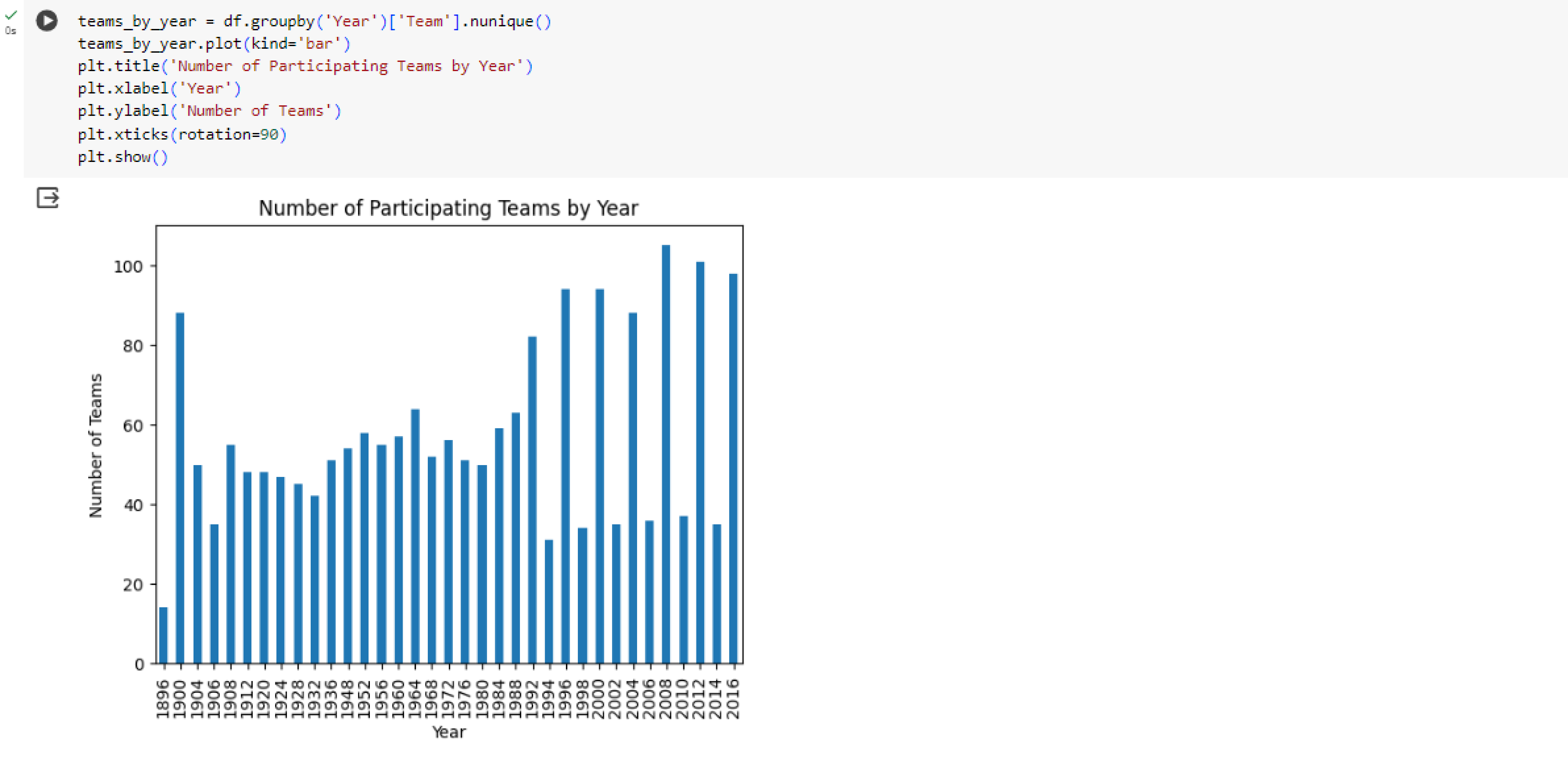


```
gold_medals_by_team = df[df['Medal'] == 'Gold'].groupby('Team').size().sort_values(ascending=False)
gold_medals_by_team.head(10).plot(kind='bar')
plt.title('Most Winning Gold Medal Teams')
plt.xlabel('Team')
plt.ylabel('Number of Gold Medals')
plt.show()
```

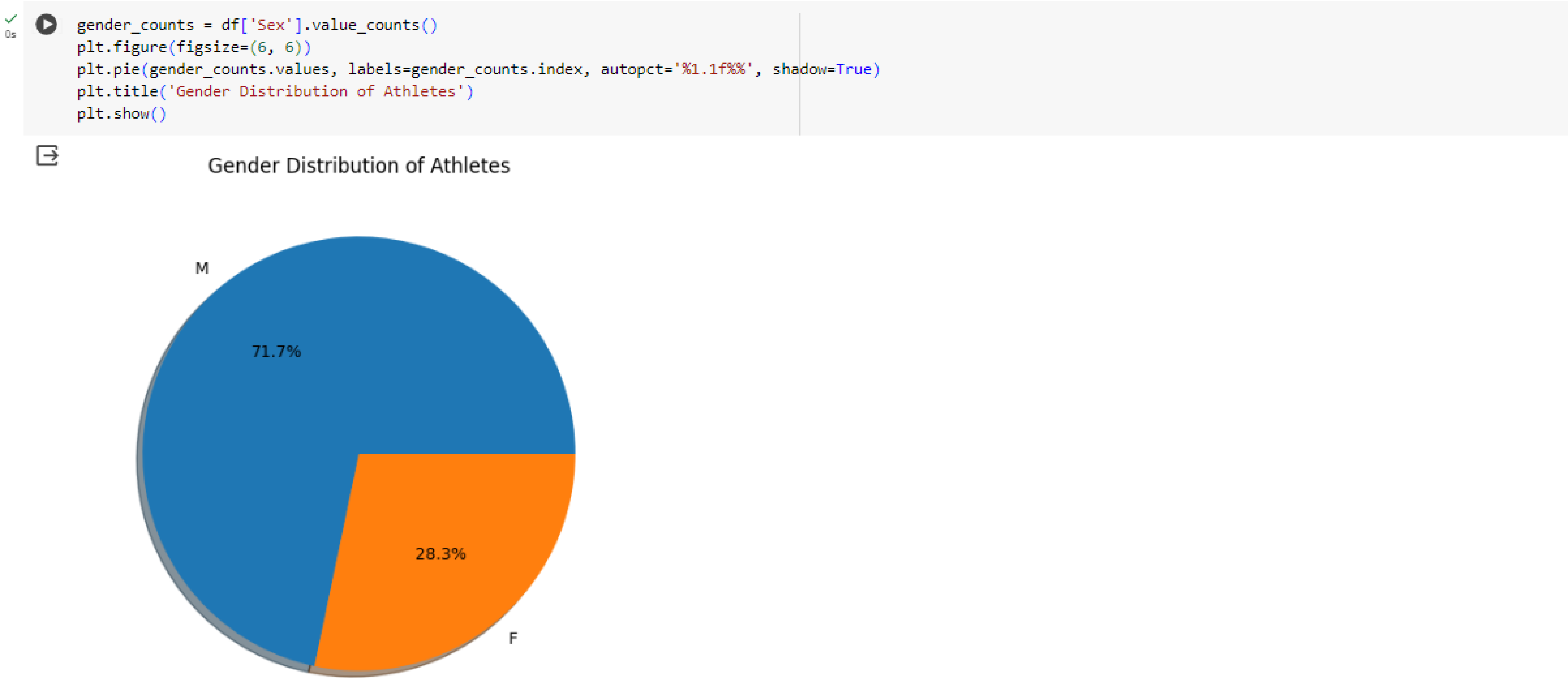
```
sports_counts = df['Sport'].value_counts()
plt.figure(figsize=(12, 6))
sns.barplot(x=sports_counts.index, y=sports_counts.values, palette='Set2')
plt.title('Distribution of Athletes by Sport')
plt.xlabel('Sport')
plt.ylabel('Number of Athletes')
plt.xticks(rotation=90)
plt.show()
```

```
<ipython-input-97-bf84540d1258>:3: FutureWarning:

Passing `palette` without assigning `hue` is deprecated and will be removed in v0.14.0. Assign the `x` variable to `hue` and set `legend=False` for the same effect.

  sns.barplot(x=sports_counts.index, y=sports_counts.values, palette='Set2')
```

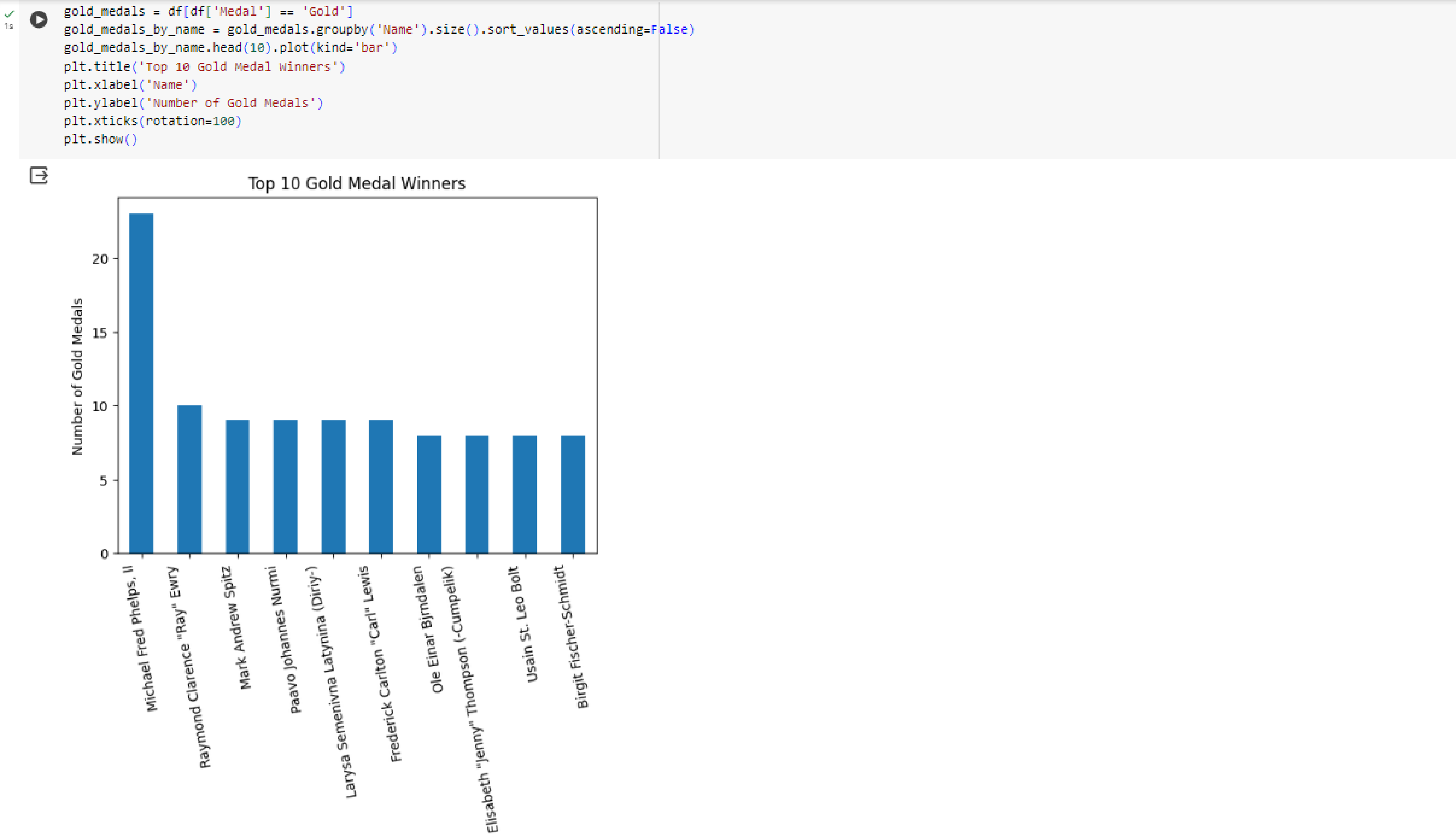




```
teams_by_year = df.groupby('Year')['Team'].nunique()
teams_by_year.plot(kind='bar')
plt.title('Number of Participating Teams by Year')
plt.xlabel('Year')
plt.ylabel('Number of Teams')
plt.xticks(rotation=90)
plt.show()
```

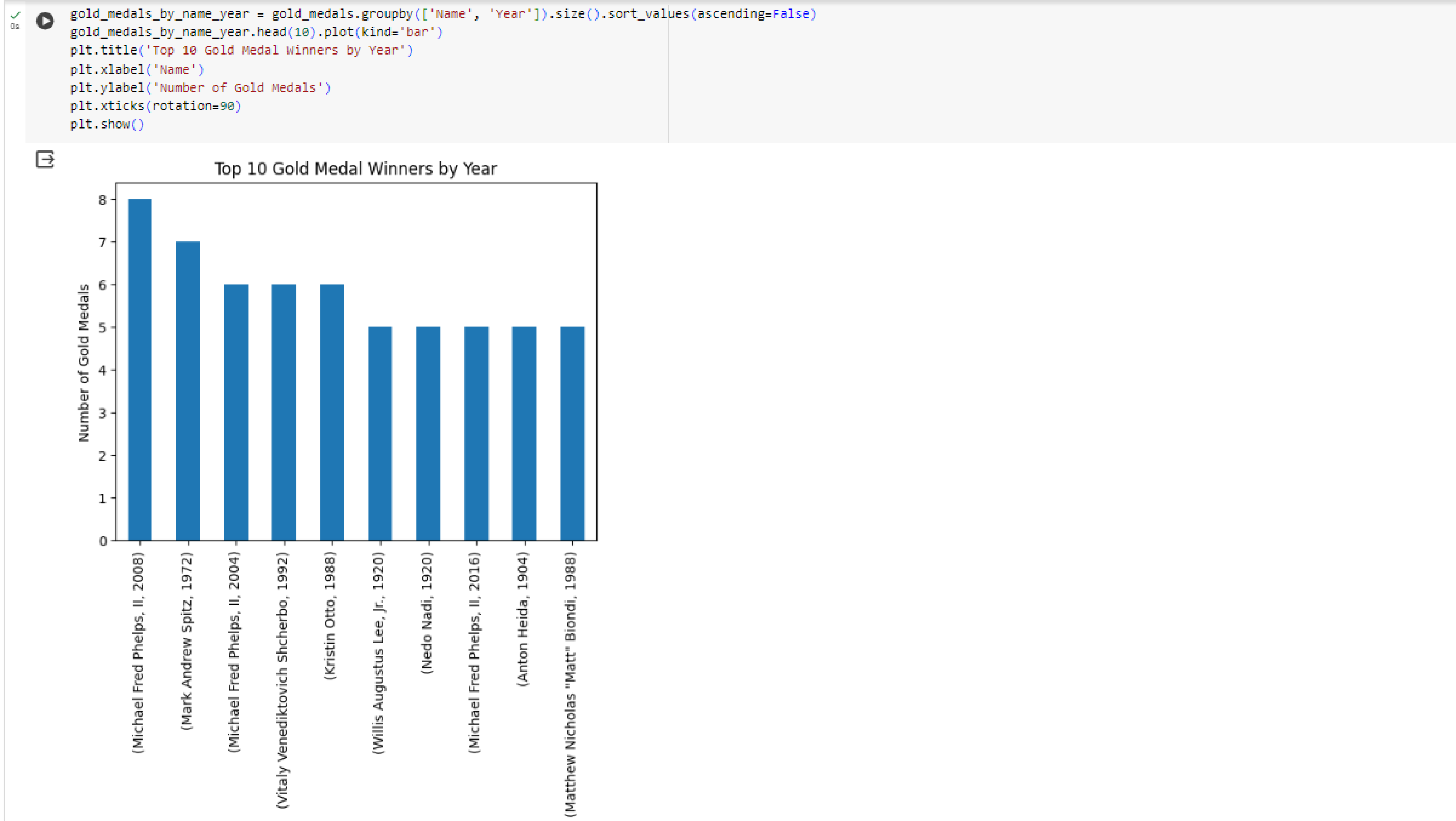


```
gender_counts = df['Sex'].value_counts()
plt.figure(figsize=(6, 6))
plt.pie(gender_counts.values, labels=gender_counts.index, autopct='%1.1f%%', shadow=True)
plt.title('Gender Distribution of Athletes')
plt.show()
```



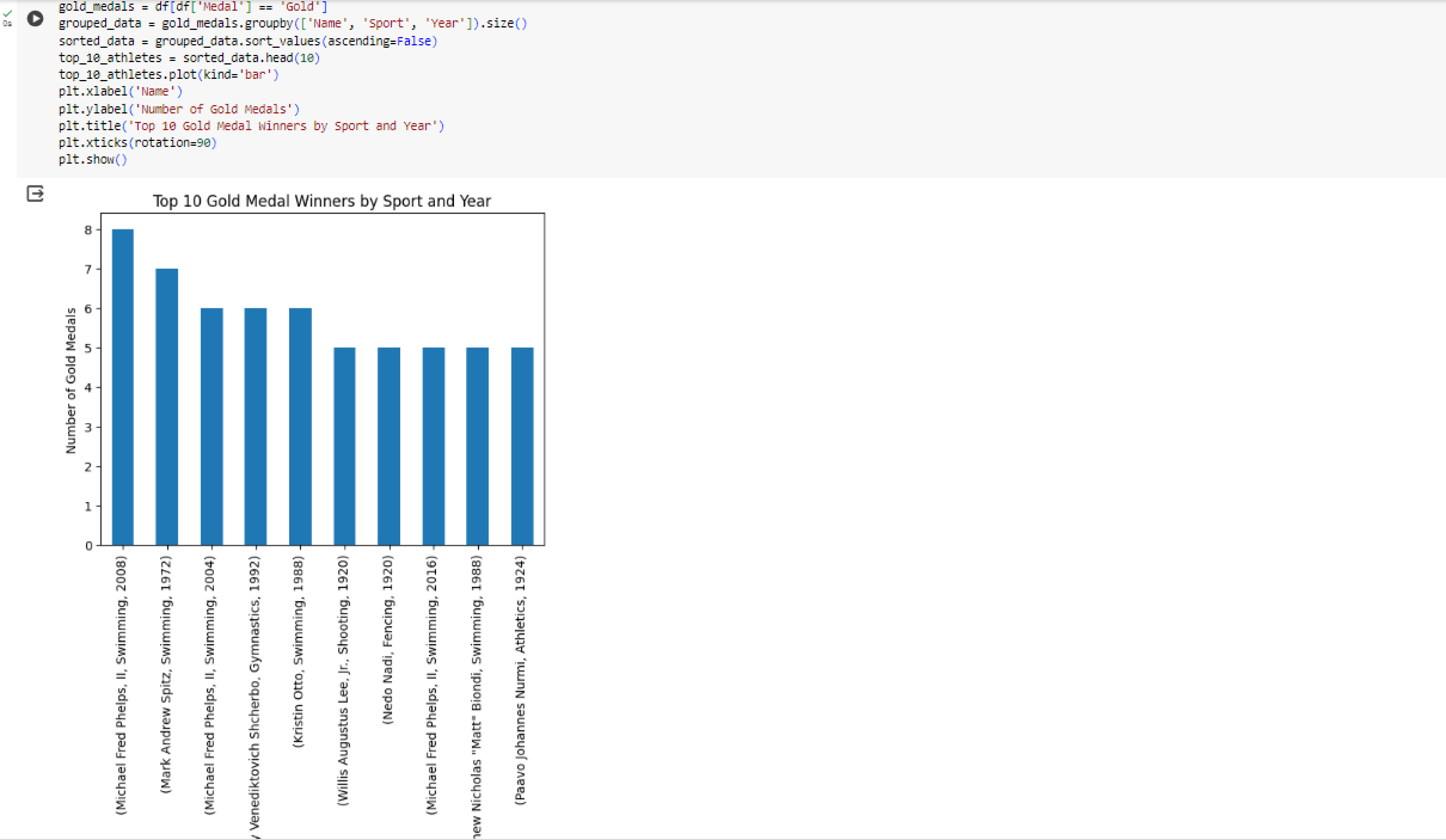
```
gold_medals = df[df['Medal'] == 'Gold']
gold_medals_by_name = gold_medals.groupby('Name').size().sort_values(ascending=False)
gold_medals_by_name.head(10).plot(kind='bar')
plt.title('Top 10 Gold Medal Winners')
plt.xlabel('Name')
plt.ylabel('Number of Gold Medals')
plt.xticks(rotation=100)
plt.show()
```

1s
Top 10 Gold Medal Winners
Number of Gold Medals
0
5
10
15
20
Michael Fred Phelps, II
Raymond Clarence "Ray" Ewry
Mark Andrew Spitz
Paavo Johannes Nurmi
Larysa Semenivna Latynina (Diriy-)
Frederick Carlton "Carl" Lewis
Ole Einar Bjrndalen
Elisabeth "Jenny" Thompson (-Cumpelik)
Usain St. Leo Bolt
Birgit Fischer-Schmidt



```
gold_medals_by_name_year = gold_medals.groupby(['Name', 'Year']).size().sort_values(ascending=False)
gold_medals_by_name_year.head(10).plot(kind='bar')
plt.title('Top 10 Gold Medal Winners by Year')
plt.xlabel('Name')
plt.ylabel('Number of Gold Medals')
plt.xticks(rotation=90)
plt.show()
```

0s
Top 10 Gold Medal Winners by Year
Number of Gold Medals
0
1
2
3
4
5
6
7
8
(Michael Fred Phelps, II, 2008)
(Mark Andrew Spitz, 1972)
(Michael Fred Phelps, II, 2004)
(Vitaly Venediktovich Shcherbo, 1992)
(Kristin Otto, 1988)
(Willis Augustus Lee, Jr., 1920)
(Nedo Nadi, 1920)
(Michael Fred Phelps, II, 2016)
(Anton Heida, 1904)
(Matthew Nicholas "Matt" Biondi, 1988)



```
gold_medals = df[df['Medal'] == 'Gold']
grouped_data = gold_medals.groupby(['Name', 'Sport', 'Year']).size()
sorted_data = grouped_data.sort_values(ascending=False)
top_10_athletes = sorted_data.head(10)
top_10_athletes.plot(kind='bar')
plt.xlabel('Name')
plt.ylabel('Number of Gold Medals')
plt.title('Top 10 Gold Medal Winners by Sport and Year')
plt.xticks(rotation=90)
plt.show()
```

Top 10 Gold Medal Winners by Sport and Year
Number of Gold Medals
0
1
2
3
4
5
6
7
8
(Michael Fred Phelps, II, Swimming, 2008)
(Mark Andrew Spitz, Swimming, 1972)
(Michael Fred Phelps, II, Swimming, 2004)
/ Venediktovich Shcherbo, Gymnastics, 1992)
(Kristin Otto, Swimming, 1988)
(Willis Augustus Lee, Jr., Shooting, 1920)
(Nedo Nadi, Fencing, 1920)
(Michael Fred Phelps, II, Swimming, 2016)
new Nicholas "Matt" Biondi, Swimming, 1988)
(Paavo Johannes Nurmi, Athletics, 1924)

Conclusion:

The analysis of the Olympic History dataset offers valuable insights into the multifaceted dynamics of the modern Olympic Games. By examining participation, performance, gender representation, and sporting diversity, the project contributes to a deeper understanding of the Olympics' evolution and its broader societal impact.