Capstone Project – Sports Stats

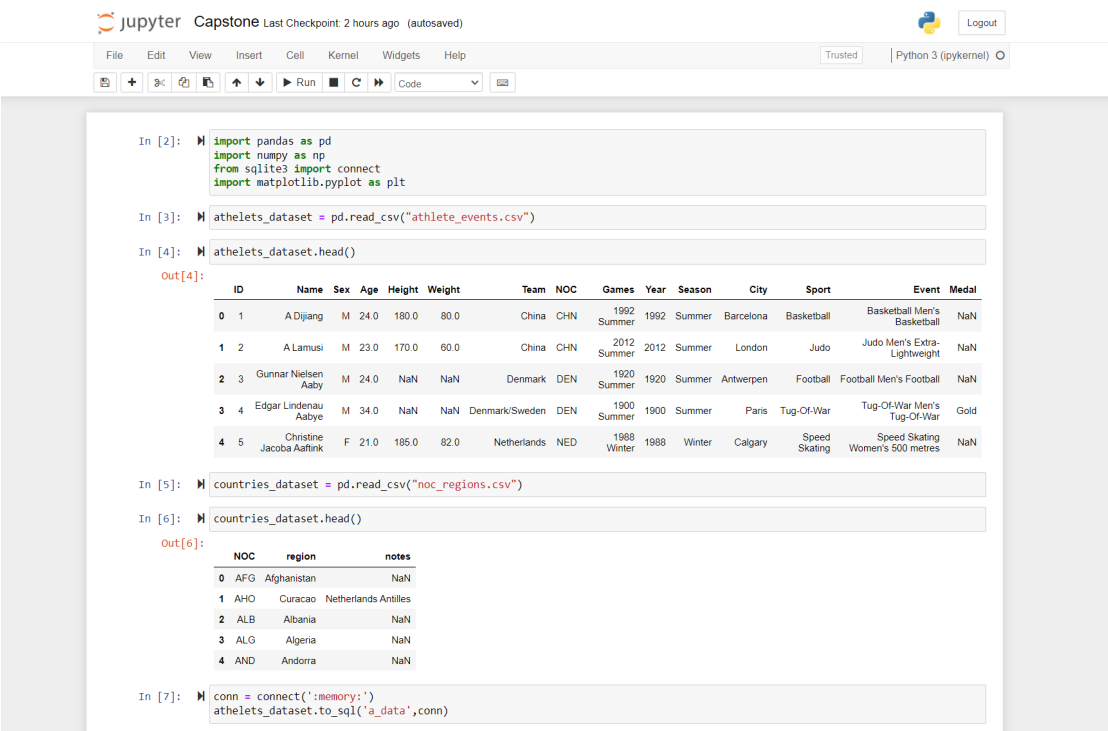
Chuxuan Quan
10/04/2021

Milestone 1

- Preparing for Your Project Proposal
  - Which client did you select and why?
    - I choose the dataset of sports stats client, because I am interested in sports and I spend much of my spare time on sports, like swimming, basketball and baseball. So I want to dig into the dataset to figure out interesting analysis.
    - Also I can find the deeper information of the sports by SQL.

  - Describe the steps you took to import and clean the data.
    - I use jupyter notebook to be my text editor and use read csv method to import data.
    - Then use pandas to sql to store the data.
    - Check the Null value and missing value by info method.

  - Perform initial exploration of data and provide some screenshots or display some stats of the data you are looking at.



	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

	NOC	region	notes
0	AFG	Afghanistan	NaN
1	AHO	Curacao	Netherlands Antilles
2	ALB	Albania	NaN
3	ALG	Algeria	NaN
4	AND	Andorra	NaN

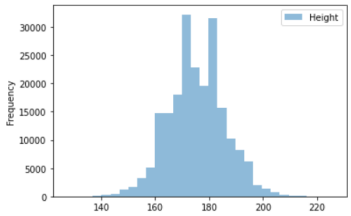
jupyter Capstone Last Checkpoint: 2 hours ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

In [11]:
```
height.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 210945 entries, 0 to 271115
Data columns (total 1 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   Height  210945 non-null  float64
dtypes: float64(1)
memory usage: 3.2 MB
```

In [12]:
```
ax = height.plot.hist(bins=30, alpha=0.5)
```



In [13]:
```
a_data = pd.read_sql("SELECT * FROM a_data",conn)
```

In [14]:
```
a_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 271116 entries, 0 to 271115
Data columns (total 16 columns):
 #   Column  Non-Null Count   Dtype
---  ------  --------------   -----
 0   index   271116 non-null  int64
 1   ID      271116 non-null  int64
 2   Name    271116 non-null  object
 3   Sex     271116 non-null  object
 4   Age     261642 non-null  float64
 5   Height  210945 non-null  float64
 6   Weight  208241 non-null  float64
 7   Team    271116 non-null  object
 8   NOC     271116 non-null  object
```

jupyter Capstone Last Checkpoint: 2 hours ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help

```
Group By Sex
""",conn)
```

In [18]:
```
gender_silver
```

Out[18]:

	Sex	Total_Silver
0	F	3735
1	M	9381

In [19]:
```
gender_bronze = pd.read_sql("""
SELECT Sex, COUNT(Medal) as Total_Bronze
  From a_data
  Where Medal = 'Bronze'
  Group By Sex
""",conn)
```

In [20]:
```
gender_bronze
```

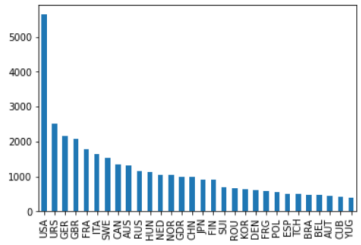
Out[20]:

	Sex	Total_Bronze
0	F	3771
1	M	9524

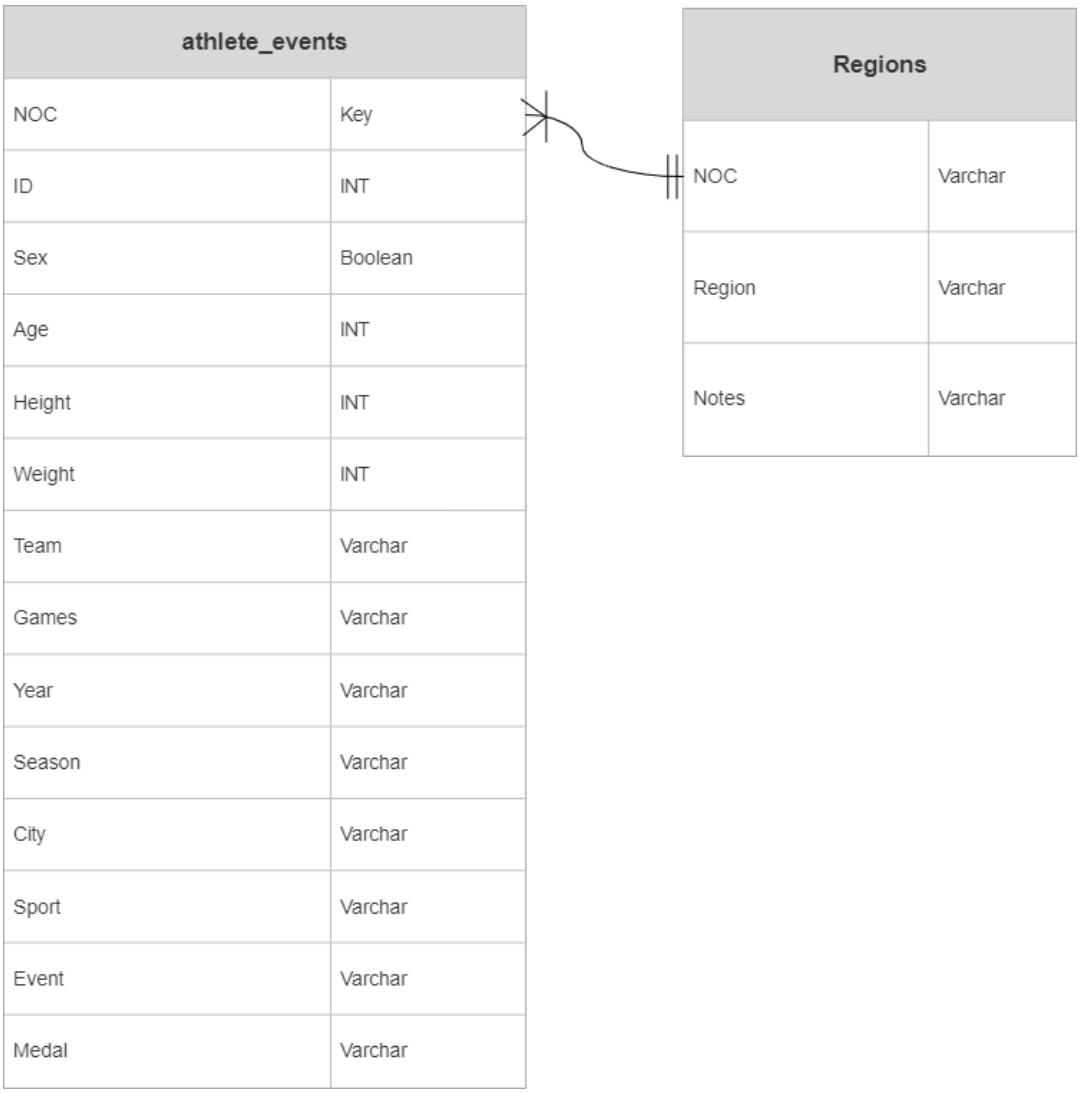
In [36]:
```
country_medal = athelets_dataset.groupby('NOC')['Medal'].count().sort_values(ascending=False)
```

In [37]:
```
country_medal_30 = country_medal[:30]
country_medal_30.plot(kind='bar')
```

Out[37]: <AxesSubplot:xlabel='NOC'>

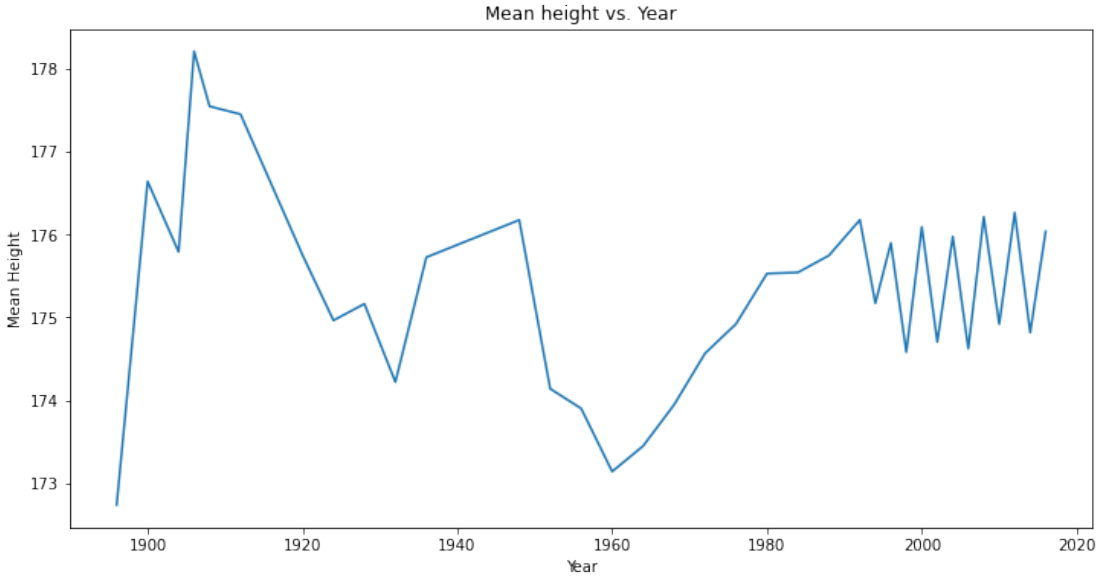


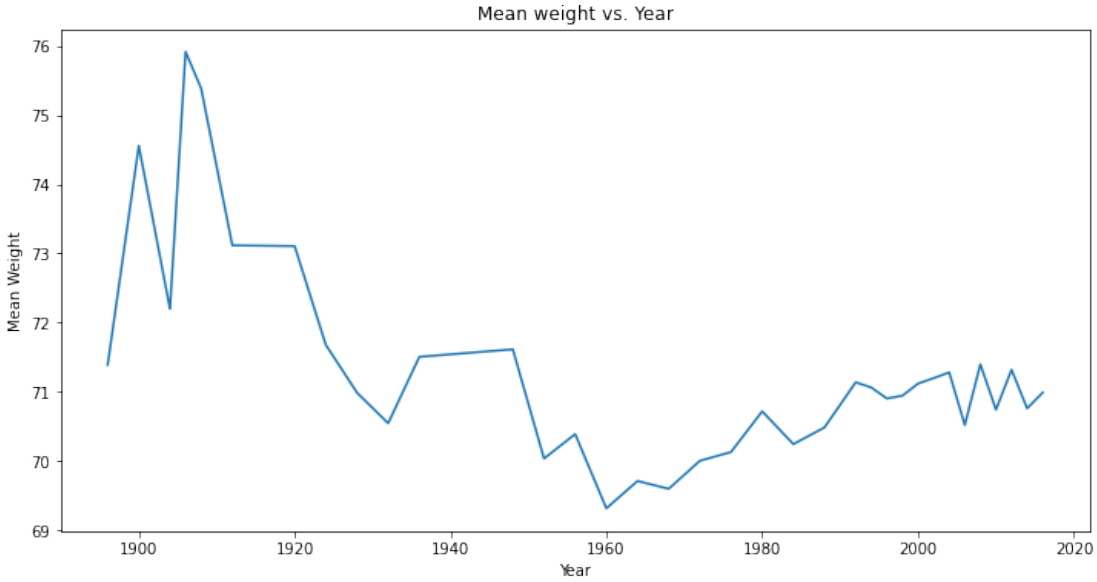
- Create an ERD or proposed ERD to show the relationships of the data you are exploring.

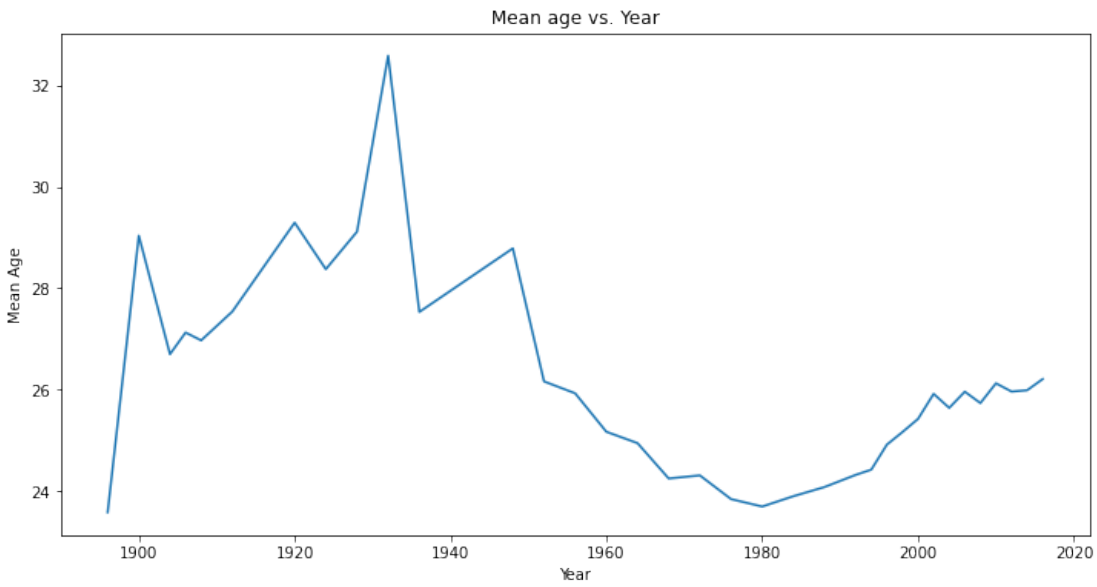


- Develop Project Proposal
  - Description
    - Analyze the trend of medals of top 30 countries with the changing of years.
    - Find the relation between different countries' performance.
    - The parameters like height, weight and age have effects on the outcome.
  - Questions
    - Is the trend of age getting younger and the trend of weight is getting larger?
    - What countries have a good performance in recent 10 years?
  - Hypothesis
    - The countries basically have no changes during the last 10 years compared it to 20 years ago.
    - The age is getting younger and the height and weight are getting larger.

Milestone 2


Mean height vs. Year
Mean Height
Year
178
177
176
175
174
173
1900
1920
1940
1960
1980
2000
2020


Mean weight vs. Year
Mean Weight
Year
76
75
74
73
72
71
70
69
1900
1920
1940
1960
1980
2000
2020



The trend of age, weight and height all experienced a rise before 1920 and a drop then a rise. Average weight and average height are highly correlated, and both reached their lowest point in 1960, and then increased year by year. The decline before 1960 may be related to the world war. In the peaceful era, the height and weight of the contestants increased. These three indicators fluctuated slightly after 2000, but they tended to stabilize as a whole.

```
year_medal_total
```
✓ 0.3s

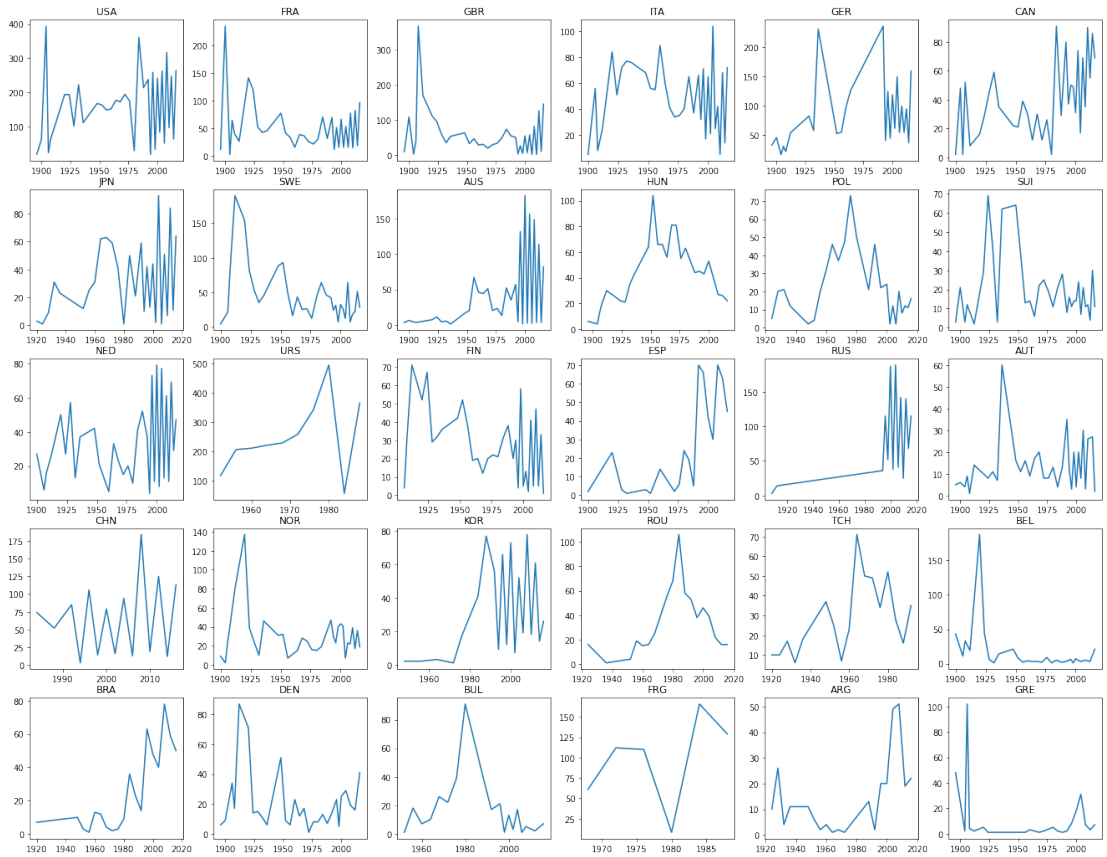
	Year	Medal
0	1896	20
1	1900	38
2	1904	59
3	1906	69
4	1908	134
5	1912	130
6	1920	182
7	1924	185
8	1928	170
9	1932	176
10	1936	193
11	1948	195
12	1952	302
13	1956	590
14	1960	976
15	1964	1176
16	1968	1248
17	1972	1401
18	1976	1498
19	1980	1572
20	1984	1683
21	1988	1827
22	1992	1834
23	1994	324
24	1996	1717
25	1998	437
26	2000	1993

```
year_medal_total.info()
```
✓ 0.1s

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 35 entries, 0 to 34
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Year    35 non-null     int64
 1   Medal   35 non-null     int64
dtypes: int64(2)
memory usage: 688.0 bytes
```

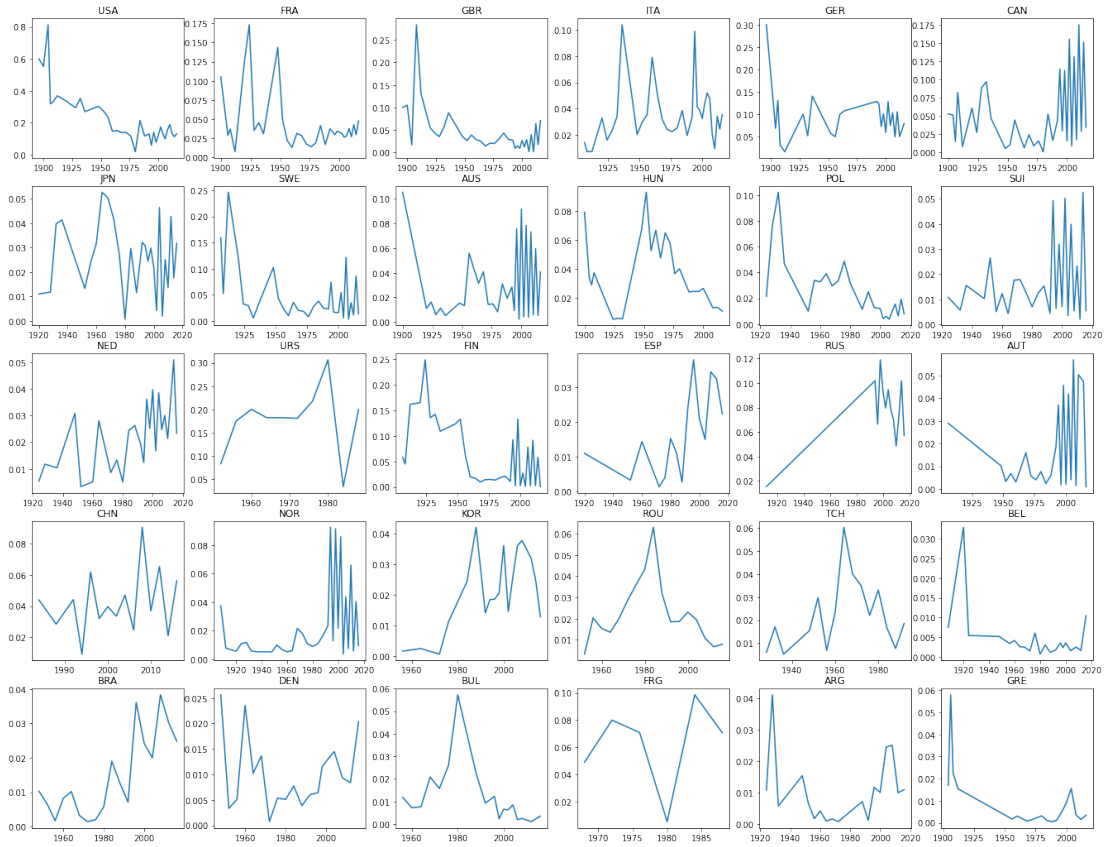
```
year_medal_total = athelets_dataset[athelets_dataset['Medal'] != 'nan']
year_medal_total = year_medal_total.groupby('Year')['Medal'].count()
year_medal_total = year_medal_total.reset_index()
```

```
plt.figure(figsize=(23,18))
for c in range(len(top_30_country)):
    p = []
    medal = athelets_dataset[(athelets_dataset['NOC'] == top_30_country[c]) & (athelets_dataset['Medal'] != 'nan')]
    medal = medal[(medal['Medal'] == 'Gold') | (medal['Medal'] == 'Silver') | (medal['Medal'] == 'Bronze')]
    medal = medal.groupby('Year')['Medal'].count()
    medal = medal.reset_index()
    plt.subplot(5,6,c+1)
    plt.plot(medal['Year'],medal['Medal'])
    plt.gca().set_title(top_30_country[c])

plt.show()
```
✓ 2.2s



Here is a list of the medals of the top 30 countries in the Olympic Games. The number of medals varies in each country. The performance of most countries is stable or on the rise, and a few countries have seen fewer and fewer medals in recent decades.

```
plt.figure(figsize=(23,18))
for c in range(len(top_30_country)):
    p = []
    medal = athelets_dataset[(athelets_dataset['NOC'] == top_30_country[c]) & (athelets_dataset['Medal'] != 'nan')]
    medal = medal[(medal['Medal'] == 'Gold') | (medal['Medal'] == 'Silver') | (medal['Medal'] == 'Bronze')]
    medal = medal.groupby('Year')['Medal'].count()
    medal = medal.reset_index()
    for i in range(len(medal['Year'])):
        p.append(medal.iloc[i,1]/int(year_medal_total[year_medal_total['Year'] == medal.iloc[i,0]]['Medal']))
    medal['Percent'] = p
    plt.subplot(5,6,c+1)
    plt.plot(medal['Year'],medal['Percent'])
    plt.gca().set_title(top_30_country[c])

plt.show()
```



This chart shows the ratio of medals of each country to total medals of each year. The first three countries, USA, FRA, GBR, decreases the ratio of medals, while CAN, NED, ESP, RUS, BRA have a growth in the medal ratio.