Athelete events data analytics

Code link:

dataset link:

❖ Objective

- Analyze trends in athletes participation across Olympic events.
- Examine key attributes such as age, height, weight, nationality, and sport.
- Identify seasonal patterns in participation and performance.
- Evaluate correlations between athlete characteristics and medal achievements.
- Provide insights into factors that may influence success at the Olympic Games.

❖ Tools and technolgies used

❖ Importing modules

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns
```

❖ Loading dataset

Code:

```
df = pd.read_csv('athlete_events_combined.csv')
df.head()
```

Output:

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	Region	Notes
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN	CHN	China
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	CHN	China
2	3	Gunnar Nielsen Aaby	M	24.0	NaN	NaN	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NaN	DEN	Denmark
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	DEN	Denmark
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	NED	Netherlands

Code :

```
df.info()
```

Output :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 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  270767 non-null  object
 16  Notes   270746 non-null  object
dtypes: float64(3), int64(2), object(12)
memory usage: 35.2+ MB
```

- Output gives the information about the data.
- Non-null count gives the count of non null rows in a column
- Dtype column shows the datatype of a columns

Task 1 : Calculating (mean, median, mode) for numerical columns Age, Height, and Weight.

Code:

```
print("Mean of Ages :", df['Age'].mean())
print("Mean of Height :", df['Height'].mean())
print("Mean of Weight :", df['Weight'].mean())
```

Output :

```
Mean of Ages : 25.556898357297374
Mean of Height : 175.33896987366376
Mean of Weight : 70.70239290053351
```

We have calculated the means of the numerical columns and the output shows the results that **Age** has mean of **25.55**, **Height** has **175.33** and the **weight** has **70.70**

Code :

```
print("Median of Ages :", df['Age'].median())
print("Median of Height :", df['Height'].median())
print("Median of Weight :", df['Weight'].median())
```

Output :

```
Median of Ages : 24.0
Median of Height : 175.0
Median of Weight : 70.0
```

Results shows the **median** of the columns, **Age** has median **24.0**, **Height** has **175.0** and the **Weight** is having the median of **70.0**.

Code:

```
print("Mode of Ages :", df['Age'].mode())
print("Mode of Height :", df['Height'].mode())
print("Mode of Weight :", df['Weight'].mode())
```

Output:

```
Mode of Ages : 0    23.0
Name: Age, dtype: float64
Mode of Height : 0    180.0
Name: Height, dtype: float64
Mode of Weight : 0    70.0
Name: Weight, dtype: float64
```

Above output shows us the **Mode value** of Age, Height and Weight column. Mode value of Age is **23.0**, for **Height** it is **180.0** and for **Weight** it is **70.0**

Task 2 : Filtering the dataset to show only athletes who participated in the 1992 Summer Olympics.
Code:

```
filtered_data = (df[(df['Season'] == 'Summer') & (df['Year'] == 1992)])
filtered_data.head()
```

Output :

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	Region	Notes
0	1	A Dijiang	M	24.0	180.0	80.0	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NaN	CHN	China
96	33	Mika Lauri Aarnikka	M	24.0	187.0	76.0	Finland	FIN	1992 Summer	1992	Summer	Barcelona	Sailing	Sailing Men's Two Person Dinghy	NaN	FIN	Finland
118	43	Morten Gjerdrum Aasen	M	34.0	185.0	75.0	Norway	NOR	1992 Summer	1992	Summer	Barcelona	Equestrianism	Equestrianism Mixed Jumping, Individual	NaN	NOR	Norway
137	50	Arvi Aavik	M	22.0	185.0	106.0	Estonia	EST	1992 Summer	1992	Summer	Barcelona	Wrestling	Wrestling Men's Heavyweight, Freestyle	NaN	EST	Estonia
160	64	M'Bairo Abakar	M	31.0	NaN	NaN	Chad	CHA	1992 Summer	1992	Summer	Barcelona	Judo	Judo Men's Half-Middleweight	NaN	CHA	Chad

Above filtered data shows the data of athletes who participated in 1992 Summer olympics.

Code :

```
print("Toal number of records athletes who participated in the 1992 Summer Olympics : ", len(filtered_data))
```

Output :

```
Toal number of records athletes who participated in the 1992 Summer Olympics :  12977
```

Total number of athletes who participated in Summer olympics 1992 is12,977.

Task 3: Count number of unique teams (NOCs) in the dataset

Code :

```
uniqe_noc = len(df['NOC'].unique())
print('Total number of unique records in NOC: ', uniqe_noc)
```

Output :

```
Total number of unique records in NOC:  230
```

Task 4: Group the data by 'Team' and calculate the average age of athletes for each team.

Code :

```
group = df.groupby('Team')['Age'].mean()
group
```

Output :

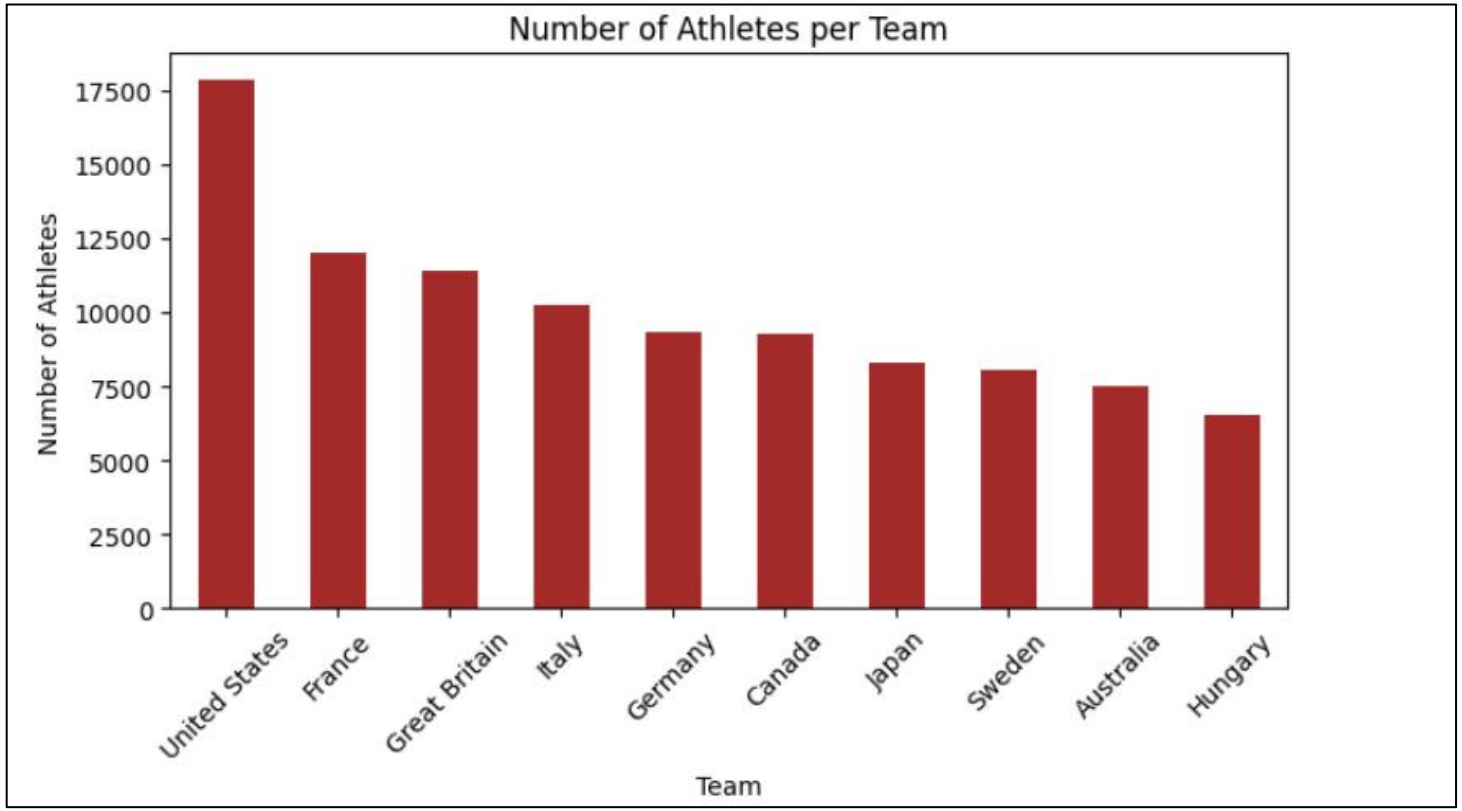
```
Team
30. Februar                33.500000
A North American Team      41.333333
Acipactli                  47.333333
Acturus                    27.000000
Afghanistan                23.538462
                               ...
Zambia                     23.461039
Zefyros                    35.500000
Zimbabwe                   25.166124
Zut                        32.000000
rn-2                       29.200000
Name: Age, Length: 1184, dtype: float64
```

- By using **groupby** function, data was grouped with same Teams.
- The other columns in the output gives the Mean of each team.

- Task 5: Create a bar plot showing the number of athletes per team.
- Code :

```
data_vis = df['Team'].value_counts()
abc = data_vis.nlargest(10, keep = 'first')
plt.figure(figsize=(8, 4))
abc.plot(kind='bar', color='Brown')
plt.title('Number of Athletes per Team')
plt.xlabel('Team')
plt.ylabel('Number of Athletes')
plt.xticks(rotation=45)
plt.show()
```

Output :



- The bar graph concludes that USA has the highest number of athletes, followed by France, Great britain, Italy.
- Australia and Hungary has the lowest number of athletes per team.

- Task 6 : Filter the data for Athletes data who won medal
- Code :

```
medal_wons = df[df['Medal'].isin(['Bronze', 'Silver', 'Gold'])]
medal_wons.head()
```

- Output :

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	Region	Notes
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	DEN	Denmark
37	15	Arvo Ossian Aaltonen	M	30.0	NaN	NaN	Finland	FIN	1920 Summer	1920	Summer	Antwerpen	Swimming	Swimming Men's 200 metres Breaststroke	Bronze	FIN	Finland
38	15	Arvo Ossian Aaltonen	M	30.0	NaN	NaN	Finland	FIN	1920 Summer	1920	Summer	Antwerpen	Swimming	Swimming Men's 400 metres Breaststroke	Bronze	FIN	Finland
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	FIN	Finland
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	FIN	Finland

Filtered data in the output is showing the first 5 rows of Athletes who won medal in any of tournament

- Task 7: Calculate Number of medals won by USA in each sport
- Code :

```
medal_wons[(medal_wons['NOC'] == 'USA')].value_counts('Sport')
```

- Output :

Sport	
Athletics	1080
Swimming	1078
Rowing	375
Basketball	341
Ice Hockey	276
Gymnastics	194
Shooting	193
Water Polo	150
Diving	140
Sailing	140
Equestrianism	132
Wrestling	128
Volleyball	120
Boxing	113
Football	102
Cycling	78
Bobsleigh	74
Speed Skating	70
Fencing	69
Baseball	68
Figure Skating	66
Tennis	62
Softball	60

Softball	60
Archery	57
Alpine Skiing	44
Weightlifting	42
Short Track Speed Skating	42
Golf	38
Rugby	36
Hockey	30
Synchronized Swimming	30
Snowboarding	24
Freestyle Skiing	21
Canoeing	21
Beach Volleyball	20
Modern Pentathlon	17
Tug-Of-War	14
Judo	14
Polo	12
Lacrosse	12
Luge	9
Taekwondo	9
Art Competitions	9
Skeleton	8
Nordic Combined	7
Curling	4
Roque	3
Triathlon	2
Ski Jumping	1
Cross Country Skiing	1
Jeu De Paume	1

Name: count, dtype: int64

- In sport Athletics and Swimming USA has won the most number of medals 1080 and 1078.
- Above 300 medals has been won by USA in Rowing and Basketball.
- Ice hockey, Gymnastics and Shooting sports has 276, 194, 193 medals.
- The sports with least medals are Ski jumping, Cross Country, Skiing having only 1 medals each.

- Task 8: Retrieve the data of Athletes winning medal in Winter olympics
- Code :

```
winter_medalists = medal_wons[(medal_wons['Season'] == 'Winter')]
winter_medalists.head()
```

- Output :

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal	Region	Notes
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	FIN	Finland
60	20	Kjetil Andr Aamodt	M	20.0	176.0	85.0	Norway	NOR	1992 Winter	1992	Winter	Albertville	Alpine Skiing	Alpine Skiing Men's Super G	Gold	NOR	Norway
61	20	Kjetil Andr Aamodt	M	20.0	176.0	85.0	Norway	NOR	1992 Winter	1992	Winter	Albertville	Alpine Skiing	Alpine Skiing Men's Giant Slalom	Bronze	NOR	Norway
63	20	Kjetil Andr Aamodt	M	22.0	176.0	85.0	Norway	NOR	1994 Winter	1994	Winter	Lillehammer	Alpine Skiing	Alpine Skiing Men's Downhill	Silver	NOR	Norway
64	20	Kjetil Andr Aamodt	M	22.0	176.0	85.0	Norway	NOR	1994 Winter	1994	Winter	Lillehammer	Alpine Skiing	Alpine Skiing Men's Super G	Bronze	NOR	Norway

Above output is showing the data of first 5 rows of Athletes who won medal in Winter olympics.

- Task 9: Filter Top 5 teams with the most gold medals.
- Code :

```
gold_medalists = df[df['Medal'] == 'Gold']

gold_medal_counts = gold_medalists['Team'].value_counts()

top_5_teams = gold_medal_counts.nlargest(5)

print(top_5_teams)
```

- Output :

```
Team
United States    2474
Soviet Union     1058
Germany           679
Italy             535
Great Britain     519
Name: count, dtype: int64
```

- Results are informing us about Top 5 teams with most number of gold medals.
- United States is leading the list with most number of gold medals in history 2474.
- Soviet union and Germany are having the count 1058 and 679 gold medals.
- Italy has 535 gold medals and Great britain on number 5 having 519 gold medals.

❖ Conclusion

- The **mean age**, **height**, and **weight** of athletes are **25.55 years**, **175.33 cm**, and **70.70 kg**, respectively.
- **Median** values closely match the means, with an age of **24 years**, height of **175 cm**, and weight of **70 kg**.
- **Mode** values indicate that the most common age is **23 years**, height is **180 cm**, and **weight** is **70 kg**.
- A total of **12,977** athletes participated in the **1992** Summer Olympics.
- The **USA** has the **highest** number of athletes and leads with **2,474 gold medals** in Olympic history.
- The Soviet Union (1,058), Germany (679), Italy (535), and Great Britain (519) follow in gold medal counts.
- Countries like **Australia** and **Hungary** have the **lowest** number of athletes per team.