DATA SCIENCE /MACHINE LEARNING PROJECT ON Tokyo 2020 Olympics Medals & Analysis

A PROJECT SUBMITTED TO BENUE STATE GOVERNMENT MINISTRY OF YOUTHS & SPORTS DEVELOPMENT

Makurdi Benue State,Nigeria

Date: 26th October,2020

BY

ADUKWU REUBEN

Tokyo 2020 Olympics Medals

DATA SCIENCE

INTRODUCTION

- It is an understatement to say that Data Science is gaining momentum. How many candidates have contacted us for Data Science opportunities? How many training programs in this field have emerged in recent years? What explains the emergence of Data Labs in organizations? All stakeholders in the ecosystem have realized the importance of this science in decision-making support and business process optimization.
- The exponential pace of data creation (Big Data), the emergence of new technologies such as Hadoop or Spark enabling the storage and processing of immense volumes of data, and the increased maturity of machine learning techniques automating the interpretation of very large databases, all contribute to the rise of Data Science.
- The growing interest in Data Science and other buzzwords such as "Machine Learning" or "Data visualization" has motivated us to shed light on this subject. This article aims to provide a concise overview of the steps that structure any Data Science project: defining objectives, data collection, data cleaning, hypothesis construction, identification of synthetic variables, building predictive models, and presenting the results.
- Data Science is the use of modern techniques and tools to analyze and extract information from data.
- In this project, we will use the Python libraries: Pandas, Numpy ,seanborn & Matplotlib.

The python libraries

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

❖ The first step is to import our data into our code environment.

❖ For that, we will use a fundamental library: Pandas. And we will download our dataset directly from its URL into our environment

- To read a CSV file located at "C:\Users\TOSHIBA\Desktop\projett\issa.csv" into a pandas DataFrame, you can use the pd.read_csv() function from the pandas library.
- Here's how you can do it:
- import pandas as pd

- # Replace the file path with the correct path to your CSV file
- file_path = "C:\\Users\\TOSHIBA\\Desktop\\projett\\issa.csv"

- # Read the CSV file into a DataFrame
- df = pd.read_csv(file_path)

- # Now you can work with the DataFrame 'df'

- In the pd.read_csv() function, you provide the file path as the first argument. If the file path contains backslashes (\), you need to use a double backslash (\\) to escape the backslashes properly
- After reading the CSV file into the DataFrame df, you can perform various data analysis, manipulation, or visualization tasks using the pandas library

Summary of the libraries

- Pandas, NumPy, Seaborn, and Matplotlib are powerful libraries in Python that are commonly used for data manipulation, analysis, and visualization. Here are some key rules or guidelines for using these libraries effectively:

Pandas:

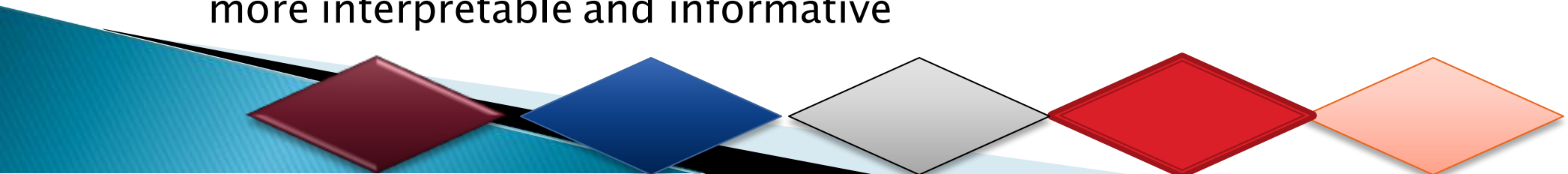
- Rule 1: Always start by importing pandas with the alias import pandas as pd for easier access to its functions and classes.
- Rule 2: Use DataFrames for tabular data and Series for one-dimensional labeled arrays.
- Rule 3: Pay attention to data types and handle missing values appropriately using functions like isnull(), fillna(), or dropna().
- Rule 4: Leverage functions like groupby(), merge(), pivot_table(), and apply() for advanced data manipulations.
- Rule 5: Utilize the powerful indexing and slicing capabilities of pandas to access and modify data efficiently.

NumPy:

- Rule 1: Import NumPy with the alias import numpy as np to use its functions and classes.
- Rule 2: Use NumPy arrays for numerical data and computations as they offer better performance compared to Python lists.
- Rule 3: Avoid using explicit loops whenever possible; instead, use NumPy's vectorized operations to perform element-wise calculations on arrays.
- Rule 4: Familiarize yourself with NumPy's mathematical functions, random number generation, and broadcasting capabilities

Seaborn:

- Rule 1: Import Seaborn with import seaborn as sns for data visualization tasks.
- Rule 2: Leverage Seaborn's high-level functions like sns.barplot(), sns.scatterplot(), sns.boxplot(), etc., for easy and informative plotting.
- Rule 3: Explore Seaborn's customization options for improving the aesthetics and information density of your plots.
- Rule 4: Always set appropriate labels, titles, and legends to make your plots more interpretable and informative



Matplotlib:

- Rule 1: Import Matplotlib with import matplotlib.pyplot as plt to create visualizations.
- Rule 2: Use plt.plot(), plt.scatter(), plt.bar(), etc., to create different types of plots.
- Rule 3: Customize your plots using plt.xlabel(), plt.ylabel(), plt.title(), and other formatting functions.
- Rule 4: Understand the different components of a Matplotlib figure, such as axes, legends, colorbars, and subplots.

General Tips:

- Practice good coding habits by commenting your code and following consistent indentation and style conventions.
- Use Jupyter Notebooks or interactive environments to explore data and visualize results step by step.
- Take advantage of the vast online documentation and tutorials available for these libraries.

By following these guidelines, you can efficiently use Pandas, NumPy, Seaborn, and Matplotlib to handle data, perform analysis, and create meaningful visualizations in your data science

```
Data=pd.read_csv(r"C:\Users\TOSHIBA\Desktop\projett\issa.csv“)
df = pd.DataFrame(Data)
df=head()
```

	Year	City	Sport	Discipline	Athlete	Country	Gender	Event	Medal
0	1896	Athens	Aquatics	Swimming	HAJOS, Alfred	HUN	Men	100M Freestyle	Gold
1	1896	Athens	Aquatics	Swimming	HERSCHMANN, Otto	AUT	Men	100M Freestyle	Silver
2	1896	Athens	Aquatics	Swimming	DRIVAS, Dimitrios	GRE	Men	100M Freestyle For Sailors	Bronze
3	1896	Athens	Aquatics	Swimming	MALOKINIS, Ioannis	GRE	Men	100M Freestyle For Sailors	Gold
4	1896	Athens	Aquatics	Swimming	CHASAPIS, Spiridon	GRE	Men	100M Freestyle For Sailors	Silver

CALCULATE THE PERCENTAGE OF MISSING DATA

- Calculate the number of missing values in each column using df.isnull().sum().
- Divide the number of missing values by the total number of rows in the DataFrame for each column.
- Multiply by 100 to get the percentage of missing data.
- Here's a Python code snippet to calculate the percentage of missing data

- import pandas as pd

- # Assuming you have already read your data into a pandas DataFrame 'df'

- # Step 1: Calculate the number of missing values in each column
- missing_values_count = df.isnull().sum()

- # Step 2: Calculate the percentage of missing data for each column
- total_rows = df.shape[0]
- percentage_missing_data = (missing_values_count / total_rows) * 100

- # Display the result
- print(percentage_missing_data)
- The code missing_values_count = df.isnull().sum() is used to calculate the number of missing values in each column of a pandas DataFrame df

- df.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 31165 entries, 0 to 31164
Data columns (total 9 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Year        31165 non-null  int64
 1   City        31165 non-null  object
 2   Sport       31165 non-null  object
 3   Discipline  31165 non-null  object
 4   Athlete     31165 non-null  object
 5   Country     31161 non-null  object
 6   Gender      31165 non-null  object
 7   Event       31165 non-null  object
 8   Medal       31165 non-null  object
dtypes: int64(1), object(8)
memory usage: 2.1+ MB
```

The output will display information about each column in the DataFrame, including the column names, non-null counts, data types, and memory usage. This summary is useful for understanding the composition and properties of the DataFrame, identifying missing or null values, and optimizing memory usage.

df.columns

Index(['Year', 'City', 'Sport', 'Discipline', 'Athlete', 'Country', 'Gender', 'Event', 'Medal'], dtype='object')

The output will be the names of all the columns in the DataFrame df. The result is displayed as a pandas Index object or a regular Python list, depending on the version of pandas being used. You can further manipulate the column names or access specific columns using the column names by indexing or slicing the DataFrame using df[column_name] or df[[column_name1, column_name2]].

Which sport is having most number of Gold Medals so far?

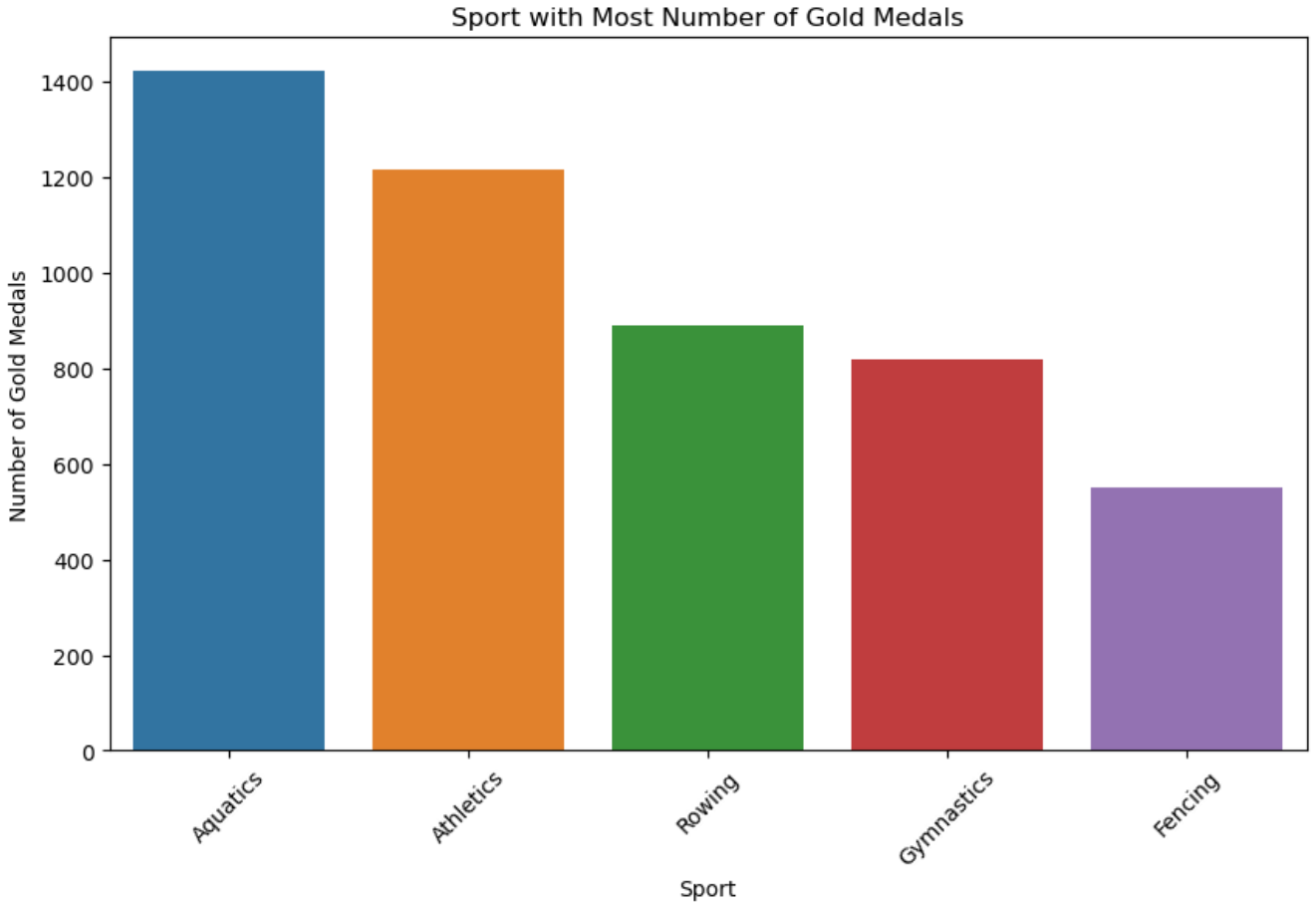
- The goal of most Olympic Games athletes is to win a gold medal, though winning any medal, even making the Olympic Games at all, is a great achievement for most of them. Here are two lists, ranking athletes by who has won the most number of medals (any color), and also by gold medals won (up to and including Tokyo 2020 results).
- The clear champion is Michael Phelps, with the most Olympic gold medals and most medals in total

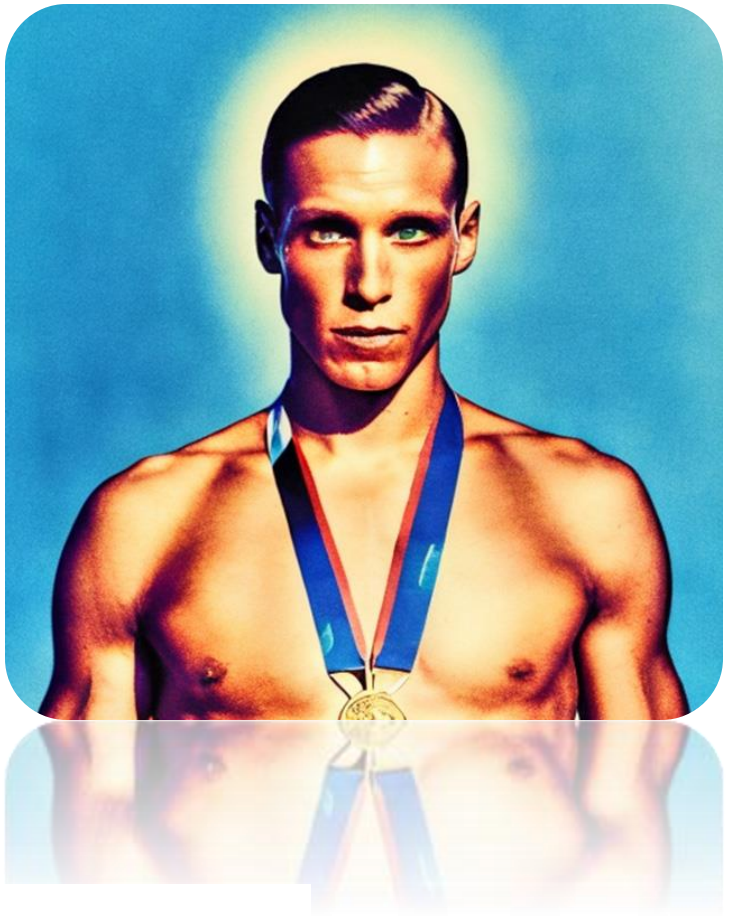
- you can use the following code

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

gold_medals = df[df['Medal'] == 'Gold']
top_sports_gold = gold_medals['Sport'].value_counts().head(5)

plt.figure(figsize=(10, 6))
sns.barplot(x=top_sports_gold.index, y=top_sports_gold.values)
plt.xlabel('Sport')
plt.ylabel('Number of Gold Medals')
plt.title('Sport with Most Number of Gold Medals')
plt.xticks(rotation=45)
plt.show()
```





The clear champion is Michael Phelps, with the most Olympic gold medals and most medals in tota

In how many cities Summer Olympics is held so far?

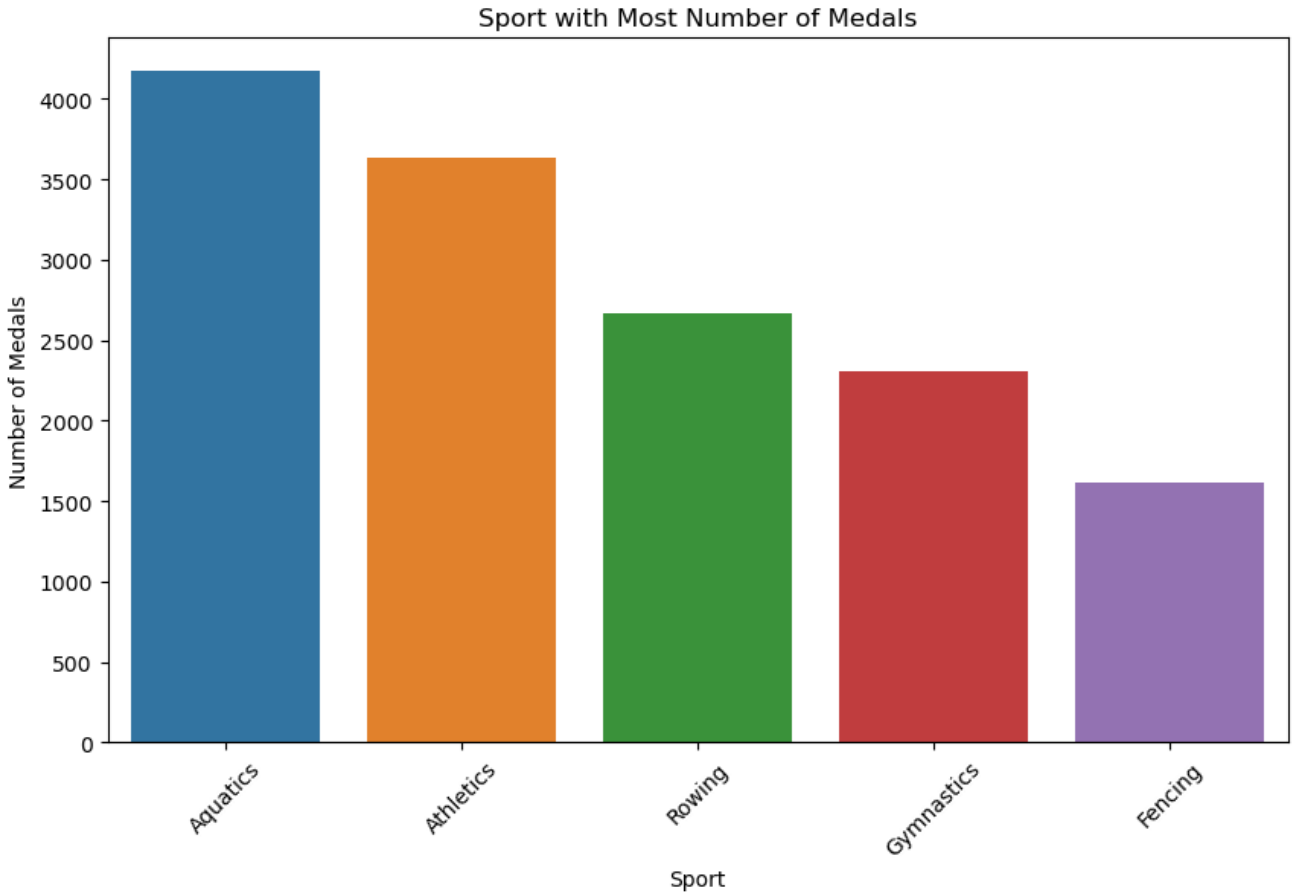
- Here is a list of all the host cities of the Summer Olympic Games since the first modern Olympics in Athens 1896. The current Olympic Games host city is Tokyo in 2021
- you can use the following code
- **num_cities = df['City'].nunique()**
- **print("Number of cities where Summer Olympics have been held:", num_cities)**

```
Number of cities where Summer Olympics have been held: 22
```

WHICH PLAYER HAS WON MOST NUMBER OF MEDALS?

- Michael Fred Phelps II is an American former competitive swimmer. He is the most successful and most decorated Olympian of all time with a total of 28 medals. Phelps also holds the all-time records for Olympic gold medals, Olympic gold medals in individual events, and Olympic medals in individual events
- you can use the following code

```
player_medals = df['Athlete'].value_counts().head()
plt.figure(figsize=(10, 6))
sns.barplot(x=player_medals.index, y=player_medals.values)
plt.xlabel('Player')
plt.ylabel('Number of Medals')
plt.title('Player with Most Number of Medals')
plt.xticks(rotation=45)
plt.show()
```



He is the most successful and most decorated Olympian of all time[7] with a total of 28 medals.

WHICH SPORT IS HAVING MOST NUMBER OF MEDALS SO FAR?

- At the Tokyo 2020 Olympic Games there were 33 sports with 50 disciplines, and a total of 339 events. This is five more sports and 18 new events compared to the previous Olympic Games. Although there are 339 medal events, because some sports award multiple bronze medalists and the occasional third place tie and other abnormalities, there are even more medals awarded.

Number of Available Gold Medals in 2021

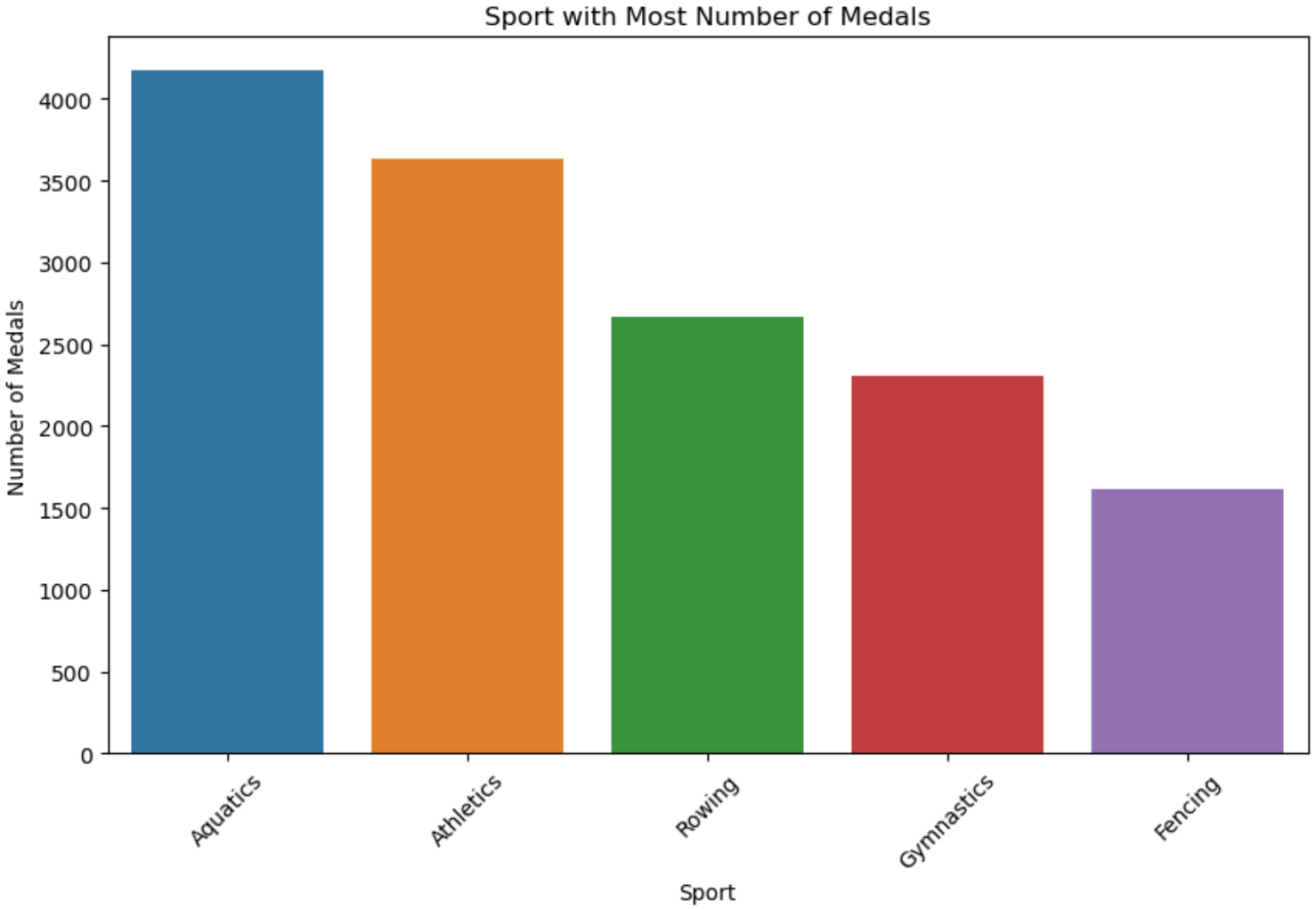
- 48 Athletics
- 37 Swimming
- 15 Judo, Shooting
- 14 Artistic Gymnastics, Rowing, Weightlifting
- 13 Boxing
- 12 Fencing, Track cycling, Canoe Sprint, Freestyle Wrestling
- 10 Sailing

- 8 Diving, Karate, Taekwondo
- 6 Equestrian, Greco-Roman Wrestling

- 5 Archery, Badminton, Table tennis, Tennis
- 4 BMX cycling, Road cycling, Skateboarding, Canoe Slalom, Basketball
- 3 Triathlon
- 2 Indoor Volleyball, Beach volleyball, Water polo, Artistic (synchronized) swimming, Field hockey, Football (Soccer), Golf, Handball, Modern pentathlon, Mountain biking, Rugby sevens, Rhythmic Gymnastics, Sport climbing, Surfing, Trampoline
- 1 Baseball, Softball

you can use the following code

- **top_sports_medals = df['Sport'].value_counts().head()**
- **plt.figure(figsize=(10, 6))**
- **sns.barplot(x=top_sports_medals.index, y=top_sports_medals.values)**
- **plt.xlabel('Sport')**
- **plt.ylabel('Number of Medals')**
- **plt.title('Sport with Most Number of Medals')**
- **plt.xticks(rotation=45)**
- **plt.show()**


Sport with Most Number of Medals
Number of Medals
0
500
1000
1500
2000
2500
3000
3500
4000
Aquatics
Athletics
Rowing
Gymnastics
Fencing
Sport

.WHICH PLAYER HAS WON MOST NUMBER GOLD MEDALS OF MEDALS?

- In cases where two or more athletes have the same number of total medals, the first tiebreaker is the number of gold medals, followed by the number of silver medals. If the tied athletes have exactly the same number of gold, silver and bronze medals, the ranking is given as a tie and the athletes are listed in order first by career years and then alphabetically by surname.
- You can follow this code

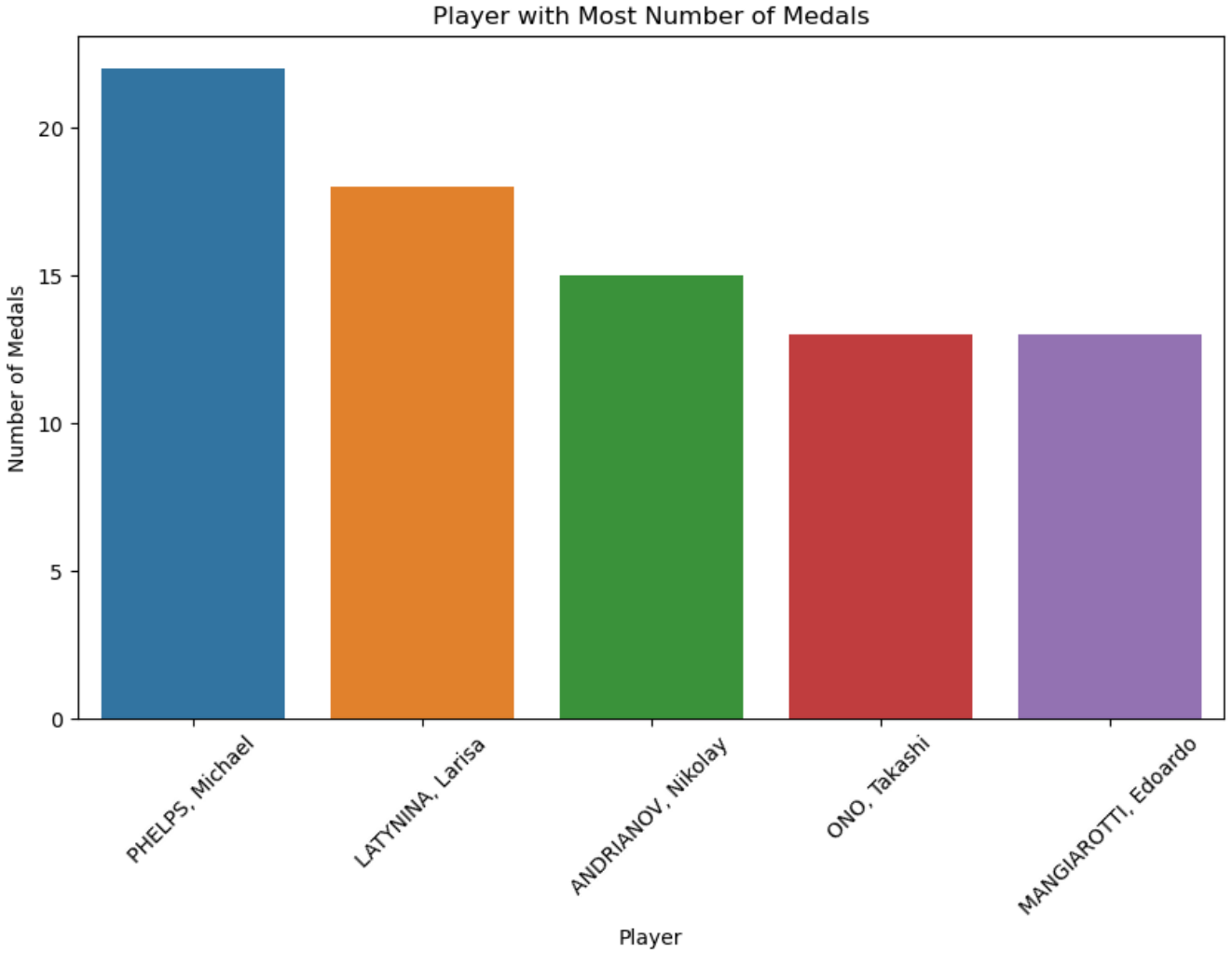
- **player_medals = df['Athlete'].value_counts().head(5)**
- **plt.figure(figsize=(10, 6))**
- **sns.barplot(x=player_medals.index, y=player_medals.values)**
- **plt.xlabel('Player')**
- **plt.ylabel('Number of Medals')**
- **plt.title('Player with Most Number of Medals')**
- **plt.xticks(rotation=45)**
- **plt.show()**

list of multiple Olympic medalists

No.	Athlete	Nation	Sport	Years	Games	Gender	Gold	Silver	Bronze	Total
1	Michael Phelps	United States	Swimming	2004–2016	Summer	M	23	3	2	28
2	Larisa Latynina	Soviet Union	Gymnastics	1956–1964	Summer	F	9	5	4	18
3	Marit Bjørgen	Norway	Cross-country skiing	2002–2018	Winter	F	8	4	3	15
4	Nikolai Andrianov	Soviet Union	Gymnastics	1972–1980	Summer	M	7	5	3	
5	Ole Einar Bjørndalen	Norway	Biathlon	1998–2014	Winter	M	8	4	1	13
6	Boris Shakhlin	Soviet Union	Gymnastics	1956–1964	Summer	M	7	4	2	
7	Edoardo Mangiarotti	Italy	Fencing	1936–1960	Summer	M	6	5	2	
	Ireen Wüst	Netherlands	Speed skating	2006–2022	Winter	F	6	5	2	
9	Takashi Ono	Japan	Gymnastics	1952–1964	Summer	M	5	4	4	
10	Paavo Nurmi	Finland	Athletics	1920–1928	Summer	M	9	3	0	

you can use the following code

- **player_medals = df['Athlete'].value_counts().head(5)**
- **plt.figure(figsize=(10, 6))**
- **sns.barplot(x=player_medals.index, y=player_medals.values)**
- **plt.xlabel('Player')**
- **plt.ylabel('Number of Medals')**
- **plt.title('Player with Most Number of Medals')**
- **plt.xticks(rotation=45)**
- **plt.show()**



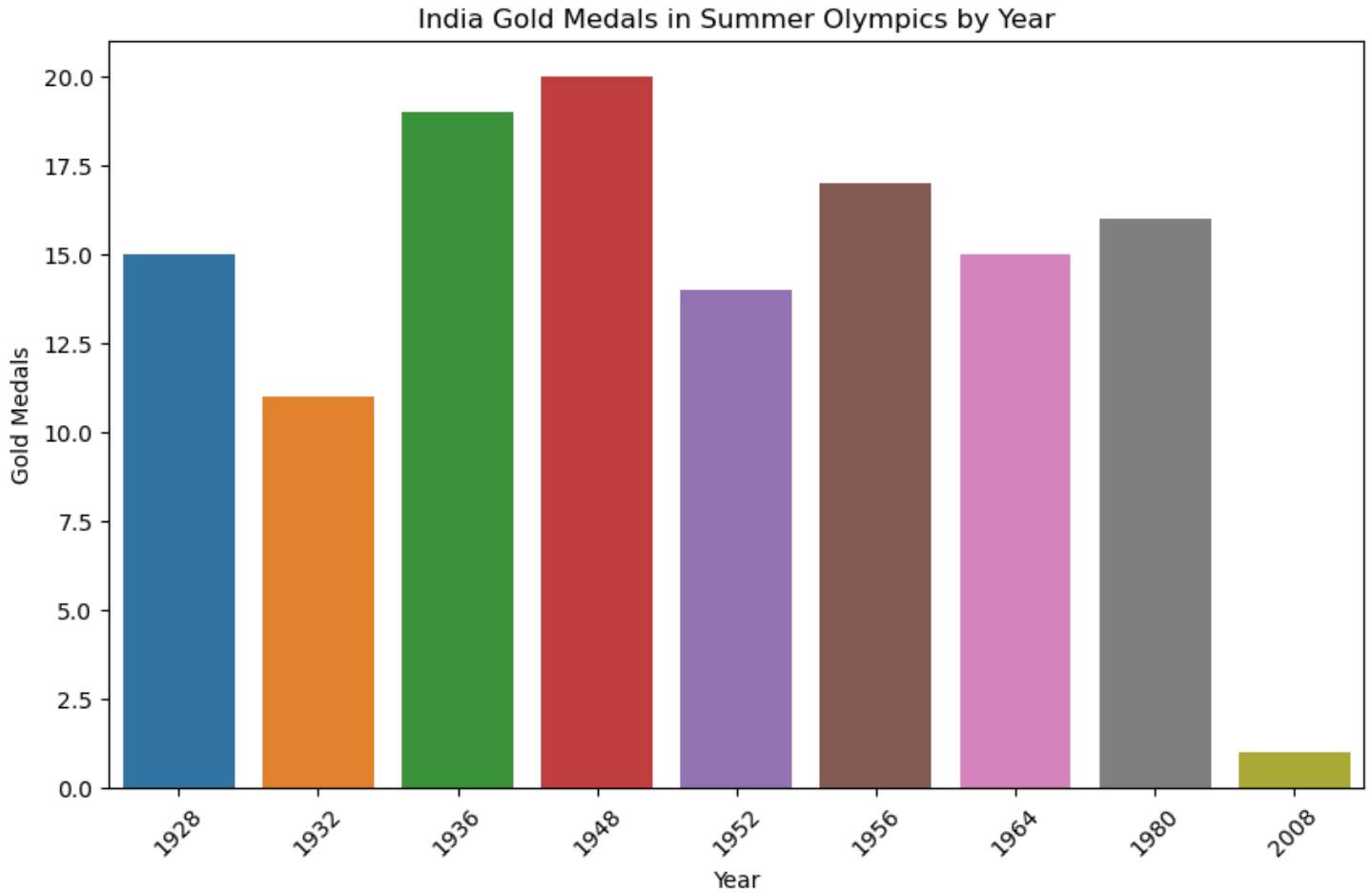
Legendary former American swimmer Michael Phelps is the most successful male Olympian of all time with 28 Olympic medals. Incredibly, 23 of them are gold medals, which is also the record for most Olympic golds won by a male athlete

In which year India won first Gold Medal in Summer Olympics?

- **India** competed at the **1948 Summer Olympics** in Wembley Park, London, England. 79 competitors, all men, took part in 39 events in 10 sports. It was the first time that India competed as an independent nation at the Olympic Games.
- You can follow this code

```
india_gold_medals = df[(df['Country'] == 'IND') & (df['Medal'] == 'Gold')]
earliest_gold_year = india_gold_medals['Year'].min()

plt.figure(figsize=(10, 6))
sns.countplot(data=india_gold_medals, x='Year')
plt.xlabel('Year')
plt.ylabel('Gold Medals')
plt.title('India Gold Medals in Summer Olympics by Year')
plt.xticks(rotation=45)
plt.show()
```



India at the 1948 Summer Olympics

IOC code	IND
NOC	Indian Olympic Association
Website	olympic.ind.in

in London

Competitors	79 (all men) in 10 sports

Medals	Gold	Silver	Bronze	Total
Ranked 22nd	1	0	0	1

Summer Olympics appearances (overview)

1900 · 1904–1912 · 1920 · 1924 · 1928 · 1932 · 1936 · **1948** · 1952 · 1956 · 1960 · 1964 · 1968 · 1972 · 1976 · 1980 · 1984 · 1988 · 1992 · 1996 · 2000 · 2004 · 2008 · 2012 · 2016 · 2020 · *2024*

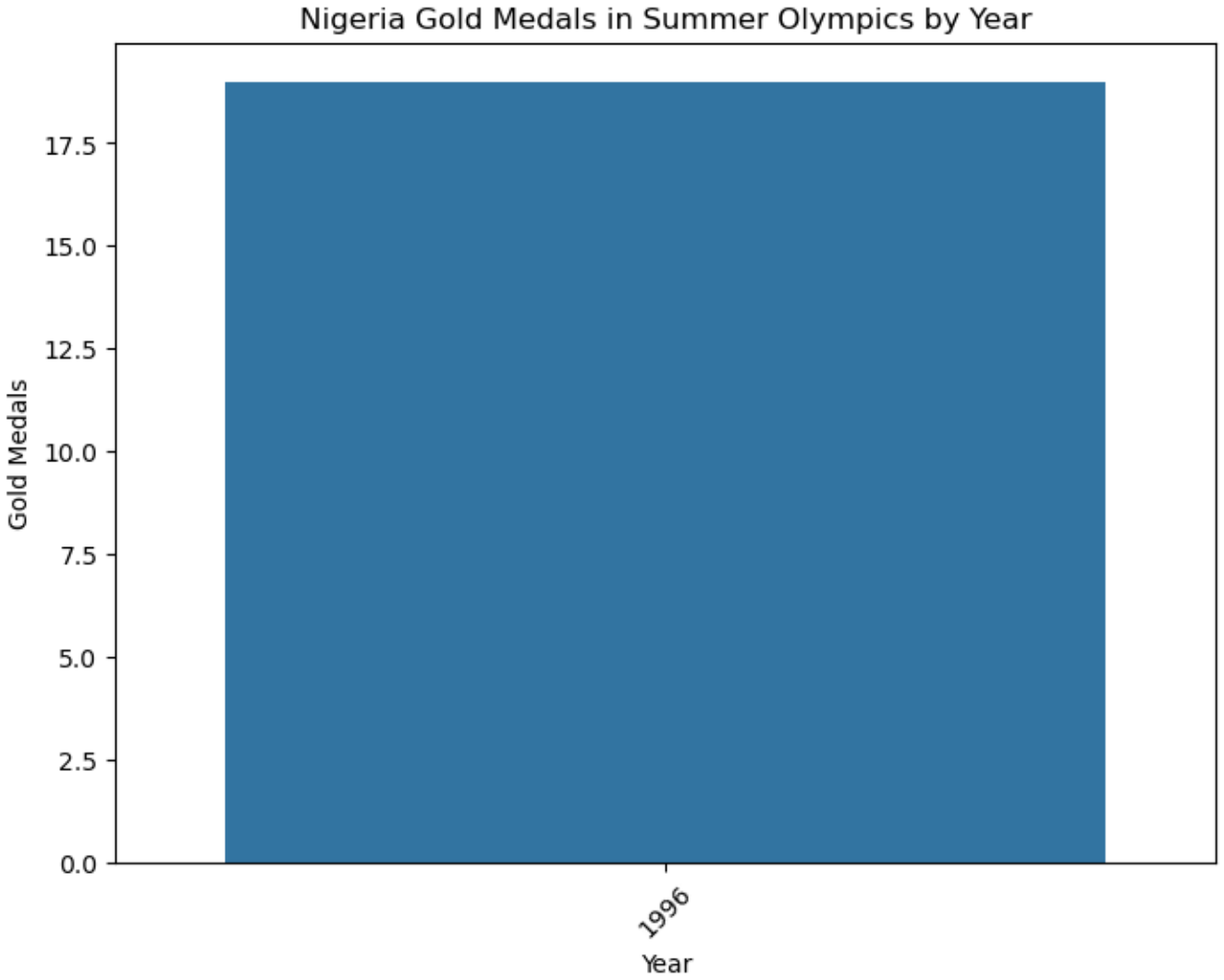
In which year Nigeria won first Gold Medal in Summer Olympics?

- Nigeria first participated in the Olympic Games in 1952, and has sent athletes to compete in every Summer Olympic Games since then, except for the boycotted 1976 Summer Olympics. The nation participated in the Winter Olympic Games in 2018, having qualified female athletes in bobsleigh and skeleton.
- Nigerian athletes have won a total of 27 medals, mostly in athletics and boxing. The national football team won the gold medal in 1996. In 2008, following the International Olympic Committee's decision to strip the American 4 × 400 metre relay team of their medals after Antonio Pettigrew confessed to using performance-enhancing drugs, their Nigerian rivals were awarded the gold medal. Nigeria also won a medal in the heavyweight division of taekwondo at the 1992 Summer Olympics; as this was only a demonstration sport, Emmanuel Oghenejobo's silver did not count as an official win

- You can follow this code
- **nigeria_gold_df = df[(df['Country'] == 'NGR') & (df['Medal'] == 'Gold')]**
- **earliest_gold_year = nigeria_gold_df['Year'].min()**

- **# Visualization**
- **plt.figure(figsize=(8, 6))**
- **sns.countplot(data=nigeria_gold_df, x='Year')**
- **plt.xlabel('Year')**
- **plt.ylabel('Gold Medals')**
- **plt.title('Nigeria Gold Medals in Summer Olympics by Year')**
- **plt.xticks(rotation=45)**
- **plt.show()**

- **print("Nigeria won its first gold medal in the Summer Olympics in", earliest_gold_year)**



1996

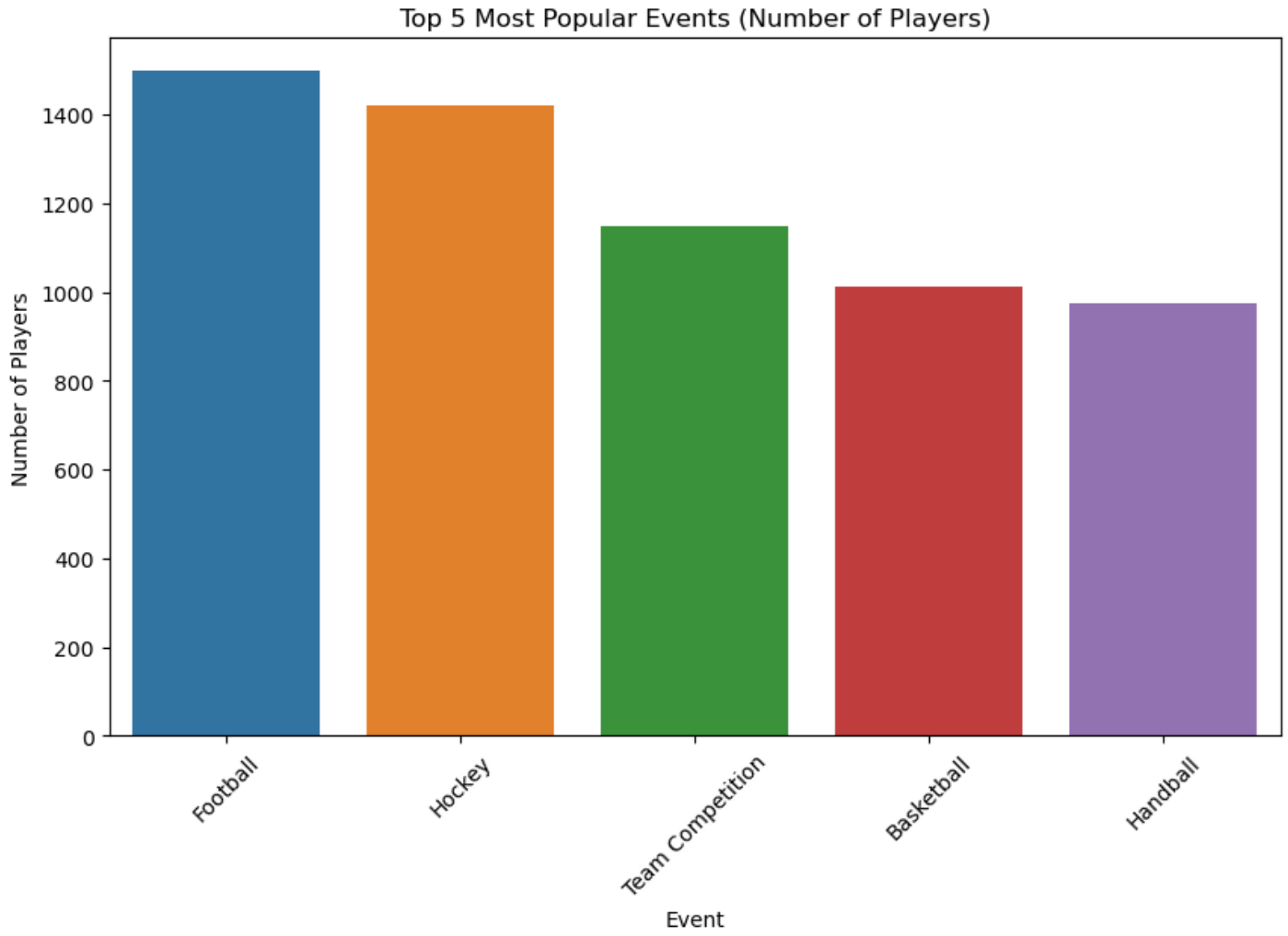
Nigeria made Olympic football history by becoming the first African and non-European and South American team to win the gold medal.

WHICH EVENT IS MOST POPULAR IN TERMS ON NUMBER OF PLAYERS? (TOP 5)

- Football is the most popular sport in the Summer Olympics. The athletes' extraordinary skill and feats, which are performed with what appears to be the greatest ease, are probably the reason why the sport is so popular. Football is so compelling to watch, which is why the Summer Olympics ratings soar during the gymnastics competitions. It combines the high-performing athleticism with the drama and pageantry inherent to the sport.

- You can follow this code
- **top_events = df['Event'].value_counts().head(5)**

- **plt.figure(figsize=(10, 6))**
- **sns.barplot(x=top_events.index, y=top_events.values)**
- **plt.xlabel('Event')**
- **plt.ylabel('Number of Players')**
- **plt.title('Top 5 Most Popular Events (Number of Players)')**
- **plt.xticks(rotation=45)**
- **plt.show()**

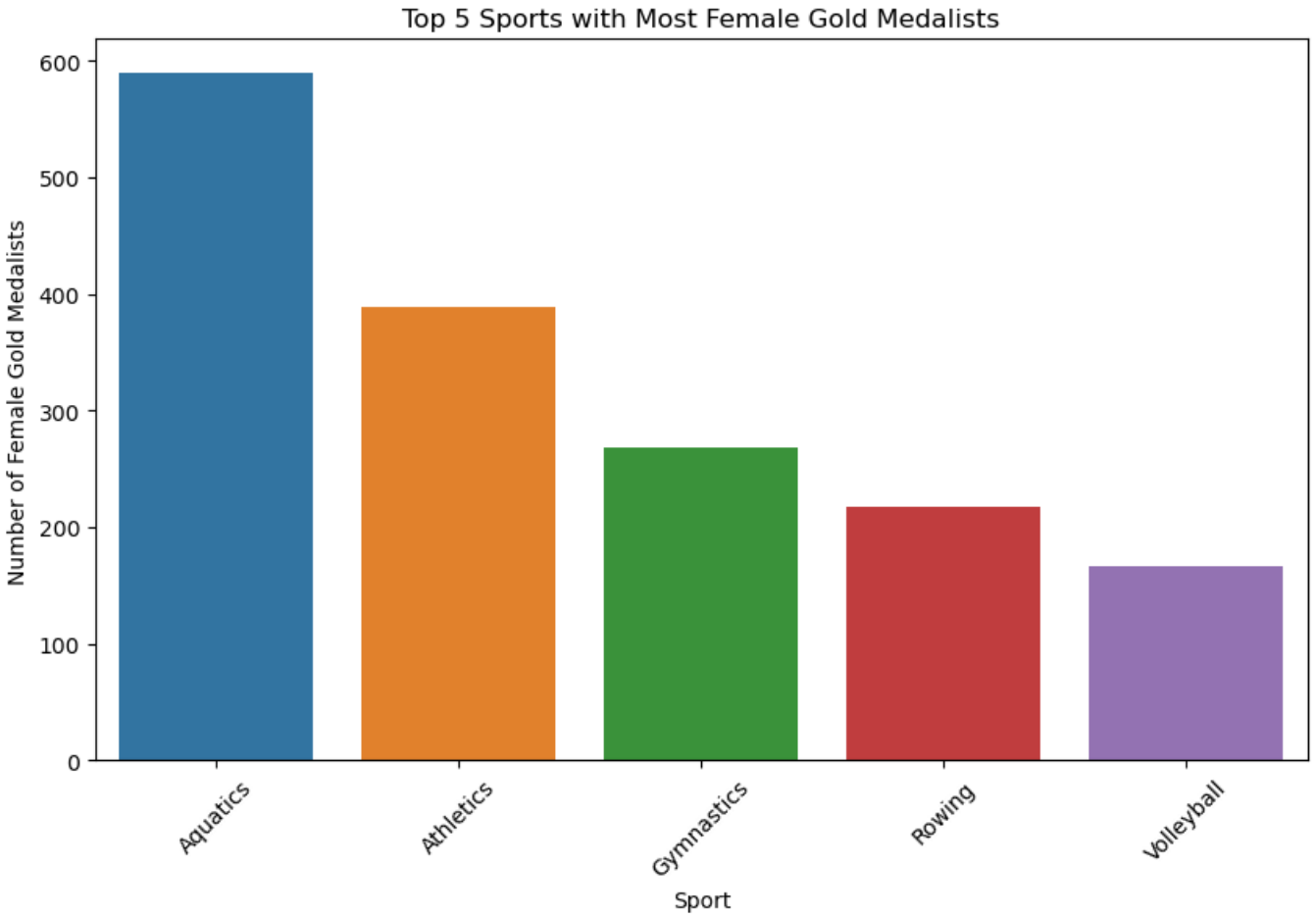

Top 5 Most Popular Events (Number of Players)
Number of Players
0
200
400
600
800
1000
1200
1400
Football
Hockey
Team Competition
Basketball
Handball
Event

WHICH SPORT IS HAVING MOST FEMALE GOLD MEDALISTS?

- After a year-long delay due to the coronavirus pandemic, the Tokyo Olympics will finally get underway from July 23. The Olympics has always been the pinnacle of sporting excellence, where some athletes attain immortal fame while a few are kept from greatness only by tantalising milliseconds and inches.
- *Sportstar takes* a look at the top male and female athletes who have won the most gold medals in the Olympics over its 31 editions since 1896.

- You can follow this code
- female_gold_medalists = df[(df['Gender'] == 'Women') & (df['Medal'] == 'Gold')]
- top_sports_female_gold = female_gold_medalists['Sport'].value_counts().head(5)

- plt.figure(figsize=(10, 6))
- sns.barplot(x=top_sports_female_gold.index, y=top_sports_female_gold.values)
- plt.xlabel('Sport')
- plt.ylabel('Number of Female Gold Medalists')
- plt.title('Top 5 Sports with Most Female Gold Medalists')
- plt.xticks(rotation=45)
- plt.show()



WOMEN

Larisa Latynina, gymnastics (URS, 1956-1964)

The most decorated female Olympic athlete of all time, Larisa Latynina, spearheaded Soviet Union's rise at the Games with 18 medals in gymnastics, nine of which were gold.

THANKS