Data Exploration and Visualisation using Python Part-1

PROJECT REPORT

TOPIC :- Tokyo Olympics

SUBMITTED BY :-
Anjali Kumari
045008

SUBMITTED TO :-
Prof. Amarnath Mitra

Project Objectives

The core objective of this project is to undertake an exhaustive analysis of the Tokyo Olympics dataset for the year 2021, scrapped from Olympics.com and https://www.bbc.com/sport/olympics/57836709. Our overarching aim is to extract valuable insights that can serve as a foundation for making informed decisions. Through a comprehensive examination of the dataset, our primary goal is to attain a profound understanding of its structural composition and the information it encompasses.

.

By applying a blend of statistical analysis and data visualization techniques, we are dedicated to discovering meaningful trends, patterns, and salient observations pertaining to athlete participation and performance. Our specific emphasis revolves around the evaluation of strategies adopted by participating nations, the identification of standout athletes representing various disciplines, and the recognition of any outstanding accomplishments.

Ultimately, our mission is to offer practical recommendations to national sports committees, sports aficionados, and other relevant stakeholders. These insights are geared toward equipping them with a holistic grasp of the Tokyo Olympics dataset, thereby empowering them to make well-informed decisions concerning athlete selection, strategic training approaches, and any pertinent considerations for the future.

General Description of Data

The dataset contains details of over 11,000 athletes, with 47 disciplines, along with 743 Teams taking part in the 2021Tokyo Olympics. This dataset contains the details of the Athletes, Coaches, Teams participating as well as the Entries by gender. It contains their names, countries represented, discipline, gender of competitors, name of the coaches.

Athletes.xlsx: Contains details about the participating Athletes [Name (name of the athlete), NOC (Country), Discipline]

```
Athletes=pd.read_excel("Athletes.xlsx")
print (Athletes)
```

```
                        Name          NOC           Discipline
0              AALERUD Katrine       Norway        Cycling Road
1                  ABAD Nestor        Spain  Artistic Gymnastics
2            ABAGNALE Giovanni        Italy              Rowing
3               ABALDE Alberto        Spain          Basketball
4                ABALDE Tamara        Spain          Basketball
...                        ...          ...                 ...
11080  ZWICKER Martin Detlef      Germany              Hockey
11081       ZWOLINSKA Klaudia       Poland        Canoe Slalom
11082            ZYKOVA Yulia          ROC            Shooting
11083       ZYUZINA Ekaterina          ROC             Sailing
11084         ZYZANSKA Sylwia       Poland             Archery

[11085 rows x 3 columns]
```

Coaches.xlsx : Contains details about the Coach(Country, Discipline, Event)

```
Coaches=pd.read_excel("Coaches.xlsx")
print (Coaches)
```

```
                    Name                         NOC         Discipline  \
0          ABDELMAGID Wael                       Egypt           Football
1               ABE Junya                       Japan         Volleyball
2            ABE Katsuhiko                       Japan         Basketball
3             ADAMA Cherif               Côte d'Ivoire           Football
4               AGEBA Yuya                       Japan         Volleyball
..                     ...                         ...                ...
389  ZAMORA PEDREIRA Javier                       Spain         Basketball
390      ZAMPIERI Francesca               Liechtenstein  Artistic Swimming
391          ZHANG Xiaohuan  People's Republic of China  Artistic Swimming
392              ZIJP Simon                 Netherlands             Hockey
393           ZONDI Nkuliso                South Africa             Hockey

     Event
0      NaN
1      NaN
2      NaN
3      NaN
4      NaN
..     ...
389    NaN
390   Duet
391    NaN
392    NaN
393  Women

[394 rows x 4 columns]
```

EntriesGender.xlsx : Contains details about the Coach(Country, Discipline, Event)

```
Gender=pd.read_excel("EntriesGender.xlsx")
print (Gender)
```

	Discipline	Female	Male	Total
0	3x3 Basketball	32	32	64
1	Archery	64	64	128
2	Artistic Gymnastics	98	98	196
3	Artistic Swimming	105	0	105
4	Athletics	969	1072	2041
5	Badminton	86	87	173
6	Baseball/Softball	90	144	234
7	Basketball	144	144	288
8	Beach Volleyball	48	48	96
9	Boxing	102	187	289
10	Canoe Slalom	41	41	82
11	Canoe Sprint	123	126	249
12	Cycling BMX Freestyle	10	9	19
13	Cycling BMX Racing	24	24	48
14	Cycling Mountain Bike	38	38	76
15	Cycling Road	70	131	201
16	Cycling Track	90	99	189
17	Diving	72	71	143

Medals.xlsx : Medals as on 29th July 2021

```
Medals=pd.read_excel("Medals.xlsx")
print (Medals)
```

```
    Rank                  Team/NOC  Gold  Silver  Bronze  Total  \
0      1   United States of America    39      41      33    113
1      2  People's Republic of China    38      32      18     88
2      3                      Japan    27      14      17     58
3      4              Great Britain    22      21      22     65
4      5                        ROC    20      28      23     71
..   ...                        ...   ...     ...     ...    ...
88    86                      Ghana     0       0       1      1
89    86                    Grenada     0       0       1      1
90    86                     Kuwait     0       0       1      1
91    86         Republic of Moldova     0       0       1      1
92    86        Syrian Arab Republic     0       0       1      1

    Rank by Total
0               1
1               2
2               5
3               4
4               3
..            ...
88             77
89             77
90             77
91             77
92             77

[93 rows x 7 columns]
```

Teams.xlsx : Contains the details of all the Teams(Country, event, Discipline, Event)

```
Teams= pd.read_excel("Teams.xlsx")
print (Teams)
```

```
                Name      Discipline                          NOC  Event
0          Belgium  3x3 Basketball                      Belgium    Men
1            China  3x3 Basketball  People's Republic of China    Men
2            China  3x3 Basketball  People's Republic of China  Women
3           France  3x3 Basketball                       France  Women
4            Italy  3x3 Basketball                        Italy  Women
..             ...             ...                          ...    ...
738   South Africa      Water Polo                South Africa  Women
739          Spain      Water Polo                       Spain    Men
740          Spain      Water Polo                       Spain  Women
741  United States      Water Polo   United States of America    Men
742  United States      Water Polo   United States of America  Women

[743 rows x 4 columns]
```

Scraped dataset- Our dataset was obtained through web scraping from two different sources, namely (https://www.bbc.com/sport/olympics/57836709 and https://olympics.com/en/olympic-games/tokyo-2020). To enhance the quality of our analysis, we subsequently partitioned this dataset into distinct subsets for more focused and effective examination.

```
from sqlalchemy.sql.expression import false
from requests.api import request
import requests
from bs4 import BeautifulSoup
import pandas as pd


# Define the URL of the website to scrape
url = "https://www.bbc.com/sport/olympics/57836709"

r= requests.get(url)
print(r)

soup= BeautifulSoup(r.text,"lxml")
table = soup.find("table",class_="gs-o-table story-body__table")
#print(table)
title = table.find_all("th")
#print(headers)
header=[]
for i in title:
  name=i.text
  header.append(name)
print(header)

df = pd.DataFrame(columns=header)
#df

rows = table.find_all("tr")
#print(rows)

for i in rows[1:]:
  data=i.find_all("td")
  row=[tr.text for tr in data]
  l=len(df)
  df.loc[l]=row
print(df)
```

Analysis

We're analyzing a team dataset, starting with fundamental checks. We first assess its shape to determine its dimensions, helping us understand its size. Then, we extract dataset information, including data types and memory usage, to comprehend its structure.

To maintain data quality, we verify the uniqueness of every value, spotting potential duplicates. Duplicate rows are also identified and handled to ensure data consistency. Crucially, we scrutinize for missing values, addressing any data gaps.

This routine process is applied to other datasets we work with, serving as a vital preprocessing step. It guarantees that we work with clean, complete, and reliable data, facilitating accurate analyses and insights.

▾ Teams_Data

```
print(Teams.head())
print("Number of rows and columns are :", Teams.shape)
print()
print( "Info about Teams :")
print(Teams.info())
print("Number of unique Team :")
print(Teams.nunique())
print("number of duplicate rows :" ,Teams.duplicated().sum())
print()
print("number of missing values : " )
print(Teams.isnull().sum())
```

```
      Name      Discipline                          NOC  Event
0  Belgium  3x3 Basketball                      Belgium    Men
1    China  3x3 Basketball  People's Republic of China    Men
2    China  3x3 Basketball  People's Republic of China  Women
3   France  3x3 Basketball                       France  Women
4    Italy  3x3 Basketball                        Italy  Women
Number of rows and columns are : (743, 4)

Info about Teams :
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 743 entries, 0 to 742
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Name        743 non-null    object
 1   Discipline  743 non-null    object
 2   NOC         743 non-null    object
 3   Event       743 non-null    object
dtypes: object(4)
memory usage: 23.3+ KB
None
Number of unique Team :
Name          146
Discipline     20
NOC            84
Event          36
dtype: int64
number of duplicate rows : 0

number of missing values :
Name          0
Discipline    0
NOC           0
Event         0
dtype: int64
```

We observe that this dataset is composed of 743 rows and 4 columns. All the variables are of object type. The number of duplicate rows is zero, and there are also no missing values

▾ Coaches_data

```
print(Coaches.head())
print("Number of rows and columns are :",Coaches.shape)
print()
print( "Info about Coaches :")
print(Coaches.info())
print("Number of unique Coaches :")
print(Coaches.nunique())
print("number of duplicate rows :" ,Coaches.duplicated().sum())
print()
print("number of missing values : " )
print(Coaches.isnull().sum())
```

```
                Name            NOC  Discipline Event
0  ABDELMAGID Wael          Egypt    Football   NaN
1        ABE Junya          Japan  Volleyball   NaN
2    ABE Katsuhiko          Japan  Basketball   NaN
3     ADAMA Cherif  Côte d'Ivoire    Football   NaN
4       AGEBA Yuya          Japan  Volleyball   NaN
Number of rows and columns are : (394, 4)

Info about Coaches :
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 394 entries, 0 to 393
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Name        394 non-null    object
 1   NOC         394 non-null    object
 2   Discipline  394 non-null    object
 3   Event       249 non-null    object
dtypes: object(4)
memory usage: 12.4+ KB
None
Number of unique Coaches :
Name          381
NOC            61
Discipline      9
Event           6
dtype: int64
number of duplicate rows : 1

number of missing values :
Name            0
NOC             0
Discipline      0
Event         145
dtype: int64
```

We observe that this dataset is composed of 394 rows and 4 columns. All the variables are of object type. We have one duplicate rows, and we have 145 missing values.

▾ Athletes_Data

```
print(Athletes.head())
print("Number of rows and columns are :",Athletes.shape)
print()
print( "Info about Athletes :")
print(Athletes.info())
print("Number of unique Athletes :")
print(Athletes.nunique())
print("number of duplicate rows :" ,Athletes.duplicated().sum())
print()
print("number of missing values : " )
print(Athletes.isnull().sum())
```

```
                Name     NOC           Discipline
0    AALERUD Katrine  Norway         Cycling Road
1        ABAD Nestor   Spain  Artistic Gymnastics
2  ABAGNALE Giovanni   Italy              Rowing
3     ABALDE Alberto   Spain           Basketball
4      ABALDE Tamara   Spain           Basketball
Number of rows and columns are : (11085, 3)

Info about Athletes :
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11085 entries, 0 to 11084
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Name        11085 non-null  object
 1   NOC         11085 non-null  object
 2   Discipline  11085 non-null  object
dtypes: object(3)
memory usage: 259.9+ KB
None
Number of unique Athletes :
Name          11062
NOC             206
Discipline       46
dtype: int64
number of duplicate rows : 1

number of missing values :
Name          0
NOC           0
Discipline    0
dtype: int64
```

We observe that this dataset is composed of 11085 rows and 3 columns. All the variables are of object type. The number of duplicate rows is one, and there are also no missing values.

EntriesGender_Data

```
print(Gender.head())
print("Number of rows and columns are :",Gender.shape)
print()
print( "Info about EntriesGender :")
print(Gender.info())
print("Number of unique EntriesGender:")
print(Gender.nunique())
print("number of duplicate rows :" ,Gender.duplicated().sum())
print()
print("number of missing values : " )
print(Gender.isnull().sum())
```

```
            Discipline  Female  Male  Total
0       3x3 Basketball      32    32     64
1              Archery      64    64    128
2  Artistic Gymnastics      98    98    196
3    Artistic Swimming     105     0    105
4            Athletics     969  1072   2041
Number of rows and columns are : (46, 4)

Info about EntriesGender :
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 46 entries, 0 to 45
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Discipline  46 non-null     object
 1   Female      46 non-null     int64
 2   Male        46 non-null     int64
 3   Total       46 non-null     int64
dtypes: int64(3), object(1)
memory usage: 1.6+ KB
None
Number of unique EntriesGender:
Discipline    46
Female        38
Male          41
Total         41
dtype: int64
number of duplicate rows : 0

number of missing values :
Discipline    0
Female        0
Male          0
Total         0
dtype: int64
```

We observe that this dataset is composed of 46 rows and 4 columns. Three variables are of int type, and one variable is of object type. The number of duplicate rows is zero, and there are also no missing values.

Medals_Data

```
print(Medals.head())
print("Number of rows and columns are :",Medals.shape)
print()
print( "Info about Medals :")
print(Medals.info())
print("Number of unique Medals:")
print(Medals.nunique())
print("number of duplicate rows :" ,Medals.duplicated().sum())
print()
print("number of missing values : " )
print(Medals.isnull().sum())
```

```
   Rank                    Team/NOC  Gold  Silver  Bronze  Total  \
0     1    United States of America    39      41      33    113
1     2  People's Republic of China    38      32      18     88
2     3                       Japan    27      14      17     58
3     4               Great Britain    22      21      22     65
4     5                         ROC    20      28      23     71

   Rank by Total
0              1
1              2
2              5
3              4
4              3
Number of rows and columns are : (93, 7)

Info about Medals :
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 93 entries, 0 to 92
Data columns (total 7 columns):
```

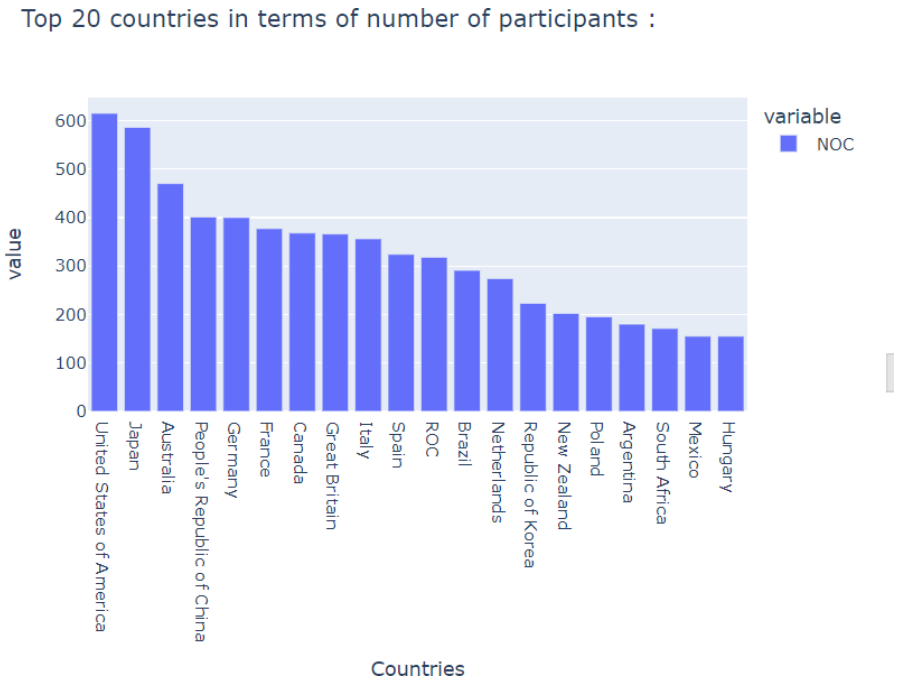
```
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Rank           93 non-null     int64
 1   Team/NOC       93 non-null     object
 2   Gold           93 non-null     int64
 3   Silver         93 non-null     int64
 4   Bronze         93 non-null     int64
 5   Total          93 non-null     int64
 6   Rank by Total  93 non-null     int64
dtypes: int64(6), object(1)
memory usage: 5.2+ KB
None
Number of unique Medals:
Rank             67
Team/NOC         93
Gold             14
Silver           17
Bronze           21
Total            30
Rank by Total    30
dtype: int64
number of duplicate rows : 0

number of missing values :
Rank             0
Team/NOC         0
Gold             0
Silver           0
Bronze           0
Total            0
Rank by Total    0
dtype: int64
```

We observe that this dataset is composed of 93 rows and 7 columns. Six variables are of int type, and one variable is of object type. The number of duplicate rows is zero, and there are also no missing values.

We'll start by tallying the number of participants from different countries and then create a bar graph to visualize this information

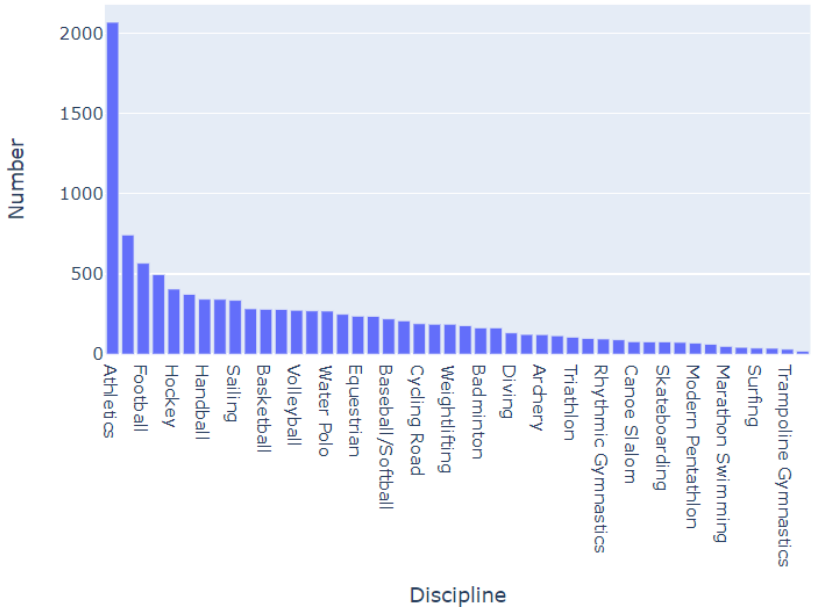
▾ Number of participants in each country

```
import plotly.express as px
data = Athletes.NOC.value_counts()
fig=px.bar(data[:20], title="Top 20 countries in terms of number of participants :")
fig.update_xaxes(title_text="Countries")
```



Next, we'll assess the count of disciplines in the Olympics. Following that, we'll examine the count of coaches from each country.

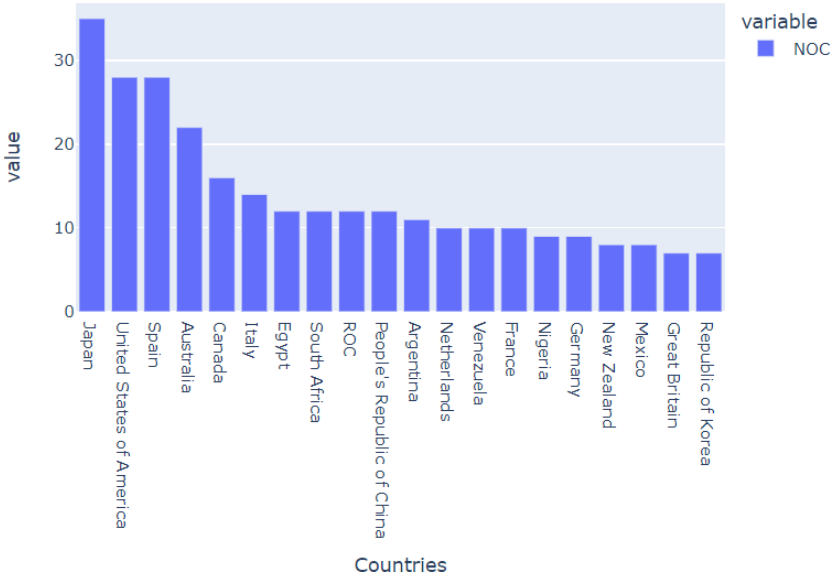
The most Discipline in the olympiad



▾ Number of Coaches in each country

```
[ ]  data2 =Coaches.NOC.value_counts()
     fig=px.bar(data2[:20], title="Top 20 countries in terms of number of Coaches :")
     fig.update_xaxes(title_text="Countries")
```

Top 20 countries in terms of number of Coaches :

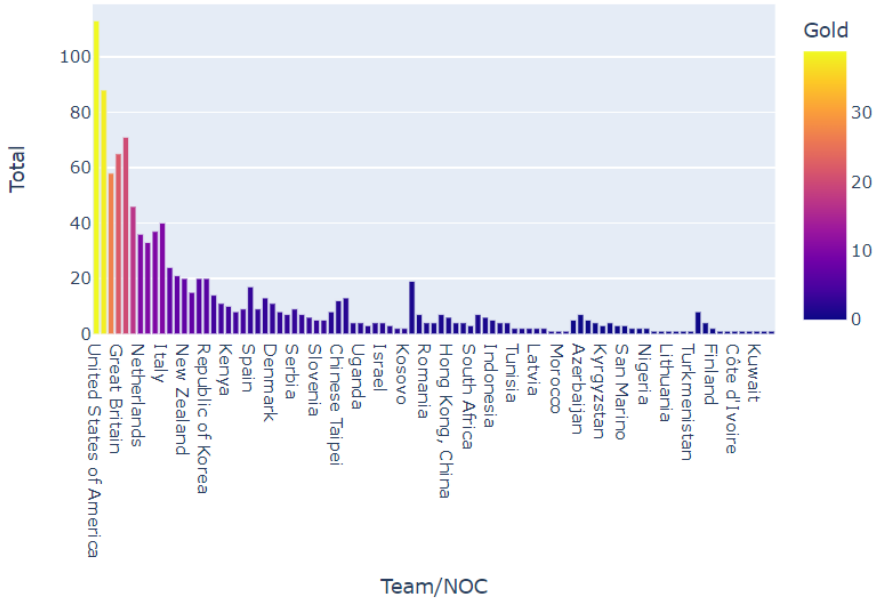


Now, we'll analyze the medal counts for each country and determine which country has received the highest number of medals.

Number of Medals in each country

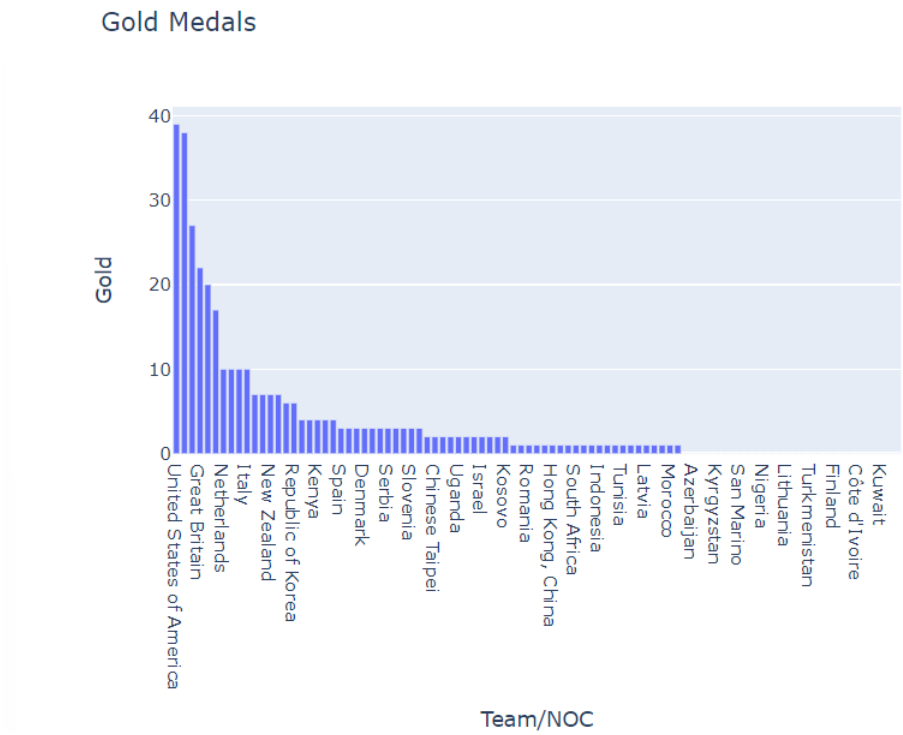
```
[ ]  px.bar(Medals, x="Team/NOC", y="Total", color="Gold", title="Top Countries in terms
```

Top Countries in terms of number of medals :

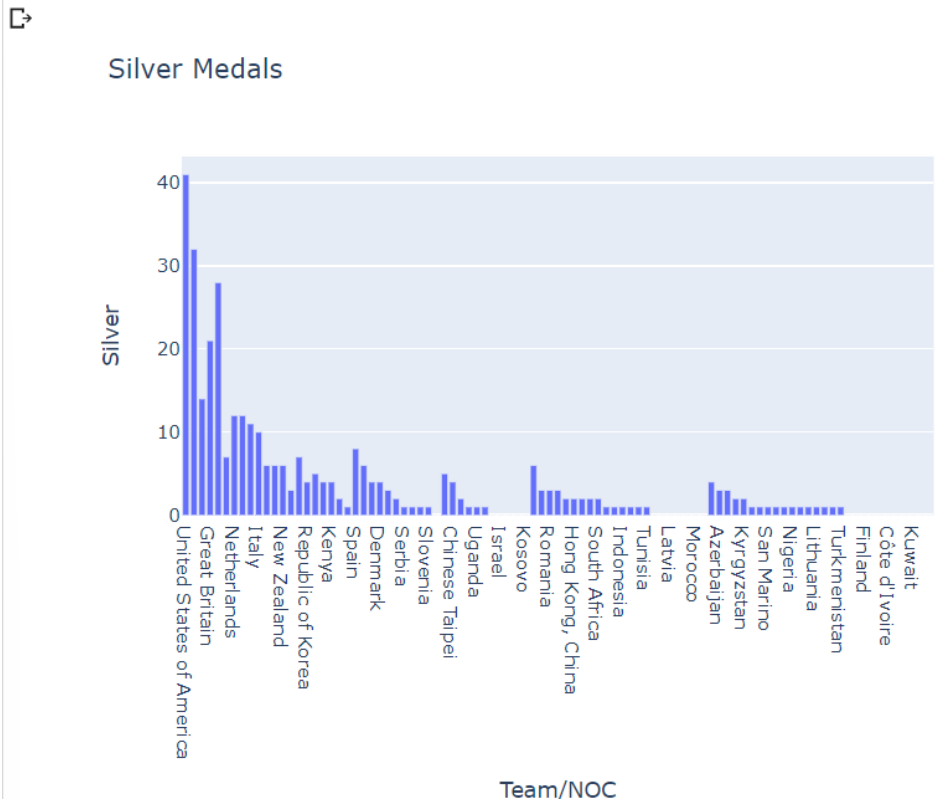


Next, we'll examine how each country has performed in terms of gold, silver, and bronze medals.

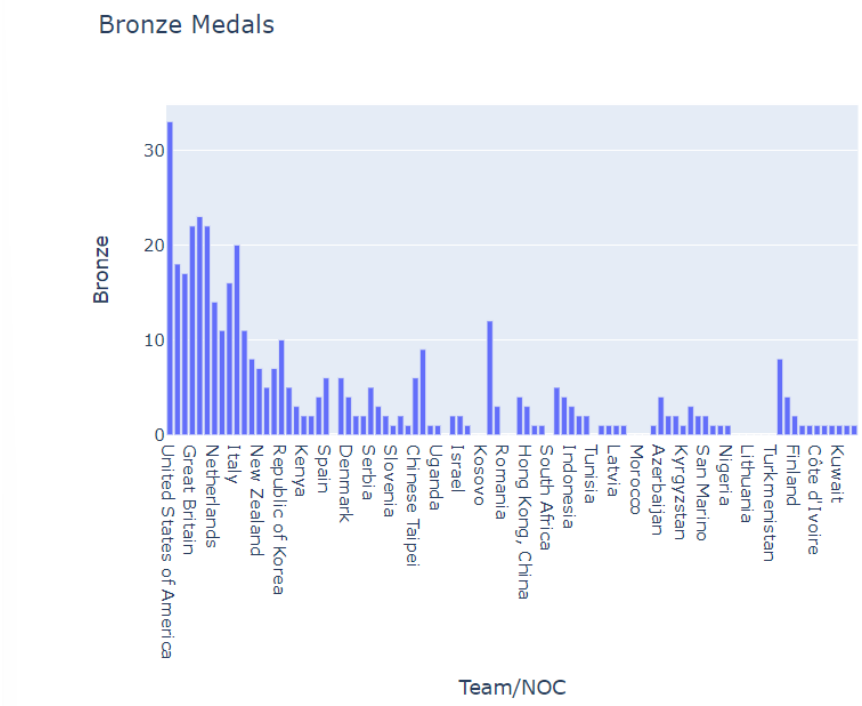
```
px.bar(Medals, x="Team/NOC", y="Gold",title="Gold Medals")
```




```
px.bar(Medals, x="Team/NOC", y="Silver" ,title="Silver Medals")
```



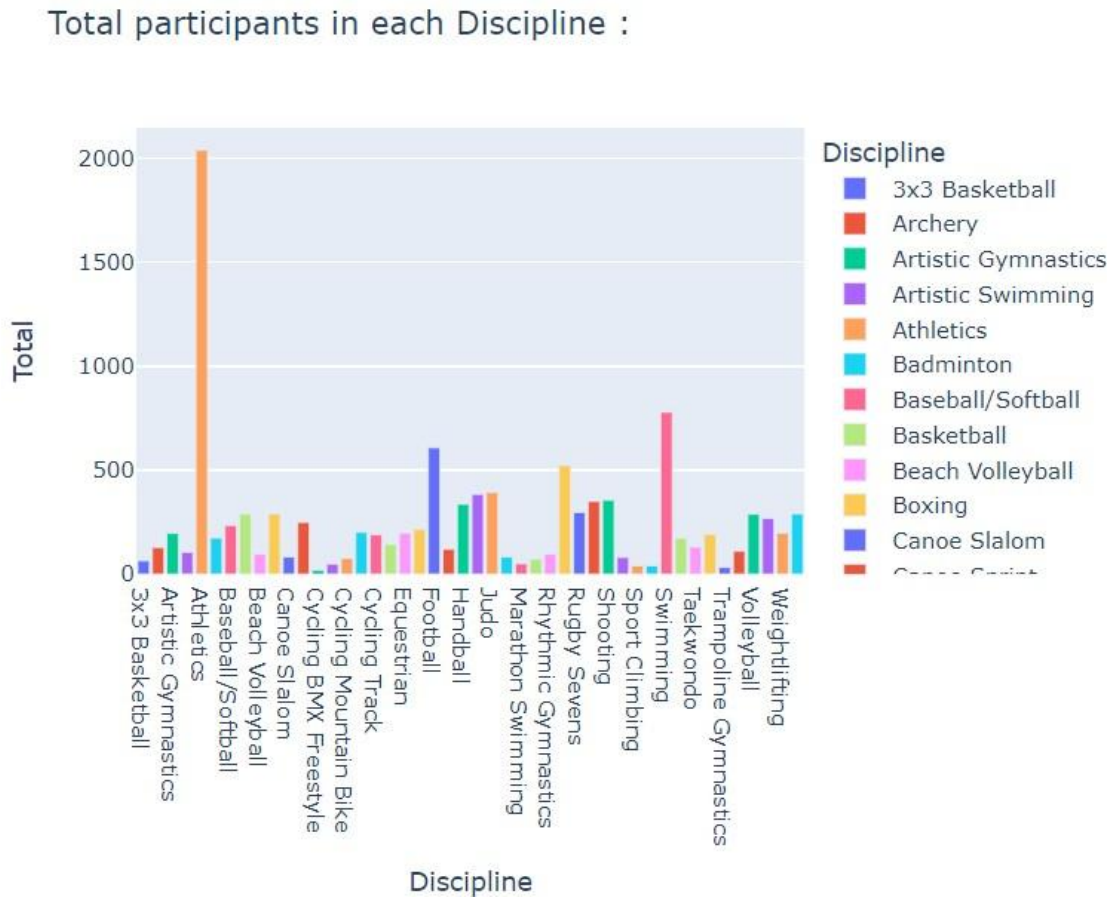
```
px.bar(Medals, x="Team/NOC", y="Bronze" ,title="Bronze Medals")
```



Now, we'll assess the total number of participants in each discipline.

▾ Number of participants in each Discipline

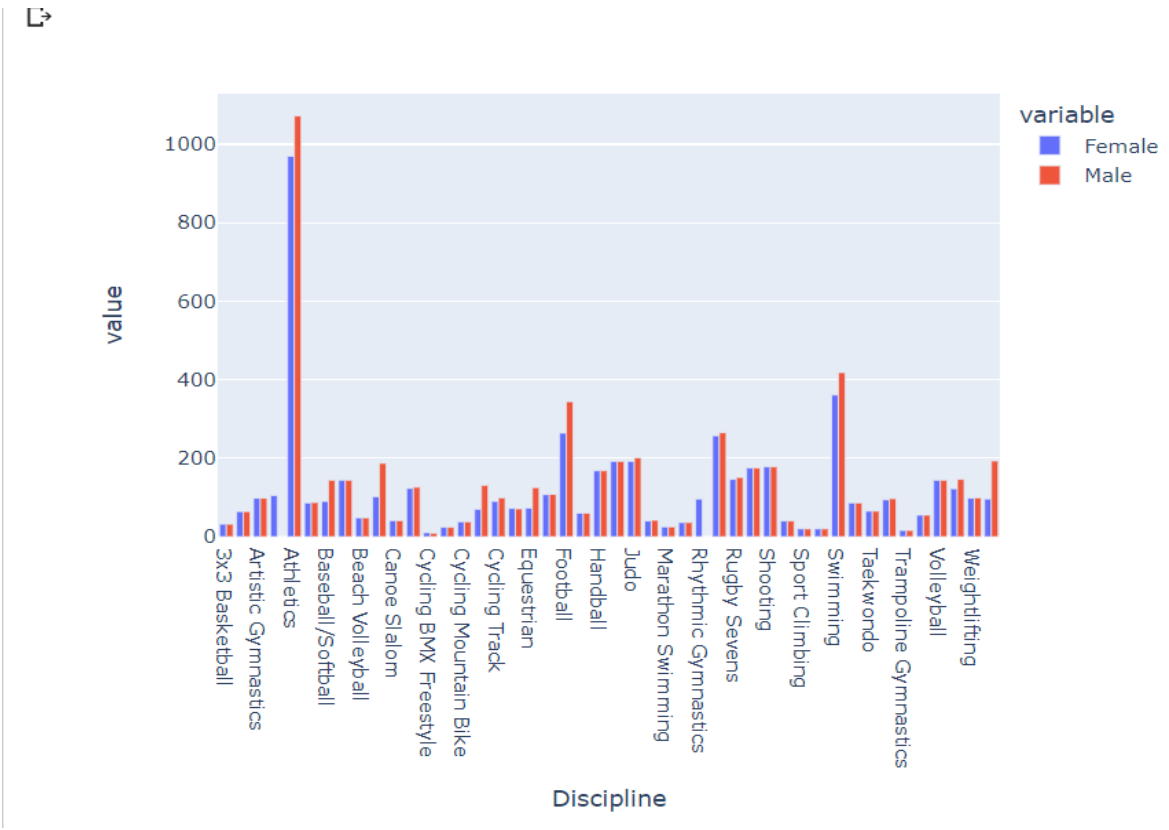
```
px.bar(Gender, x="Discipline", y="Total", color="Discipline", title="Total participants in each Discipline : ")
```



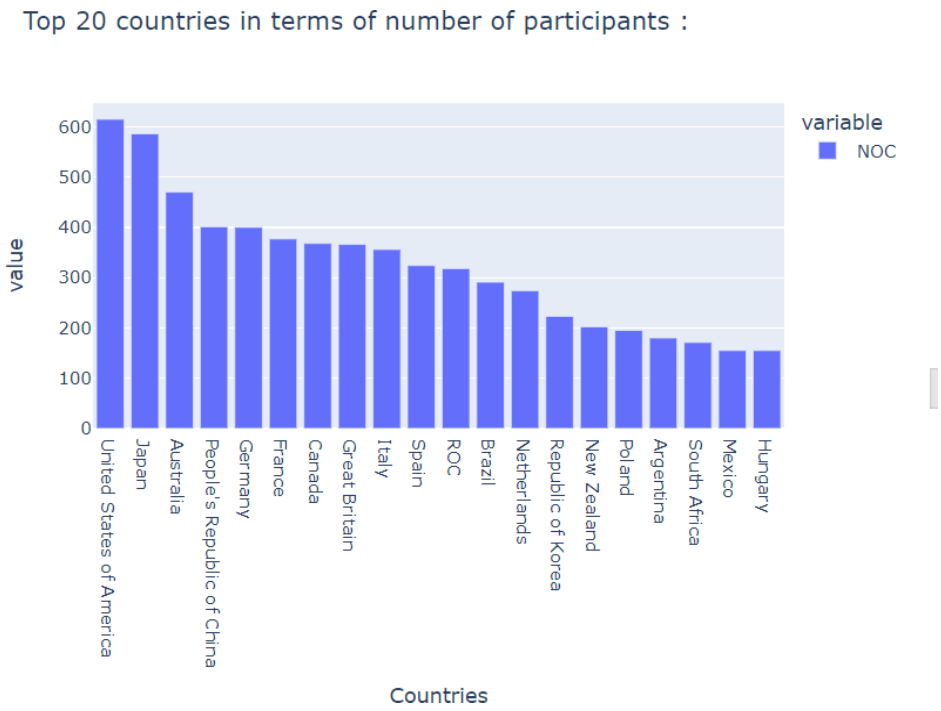
Now, we'll analyse the gender distribution, specifically the count of males and females in each discipline.

▾ Number of Male and Female in each Discipline

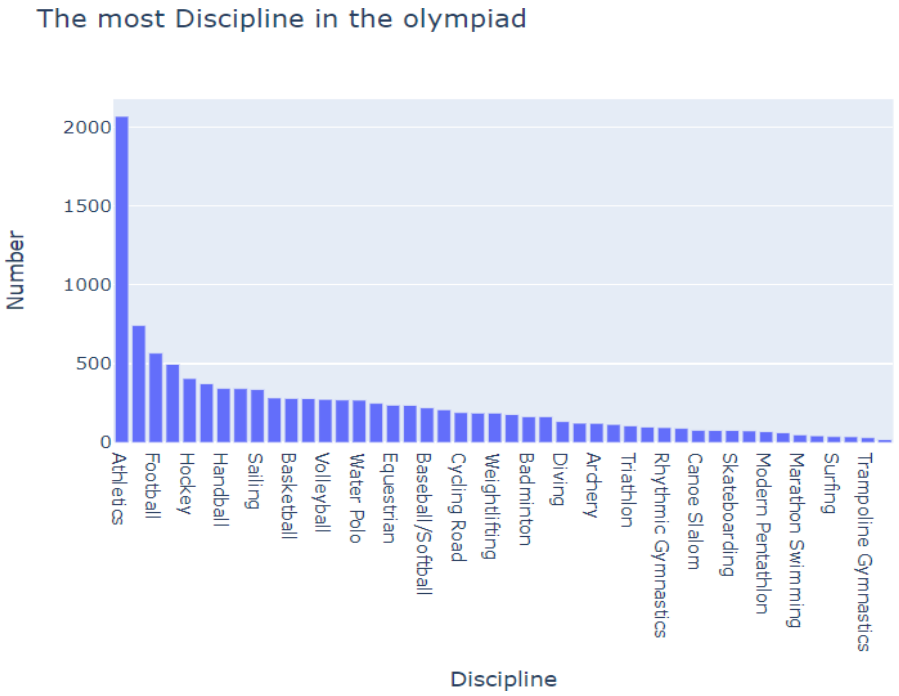
```
[ ]  px.bar(Gender, x="Discipline", y=["Female", "Male"], barmode="group"
                )
```


variable
Female
Male
value
1000
800
600
400
200
0
3x3 Basketball
Artistic Gymnastics
Athletics
Baseball/Softball
Beach Volleyball
Canoe Slalom
Cycling BMX Freestyle
Cycling Mountain Bike
Cycling Track
Equestrian
Football
Handball
Judo
Marathon Swimming
Rhythmic Gymnastics
Rugby Sevens
Shooting
Sport Climbing
Swimming
Taekwondo
Trampoline Gymnastics
Volleyball
Weightlifting
Discipline

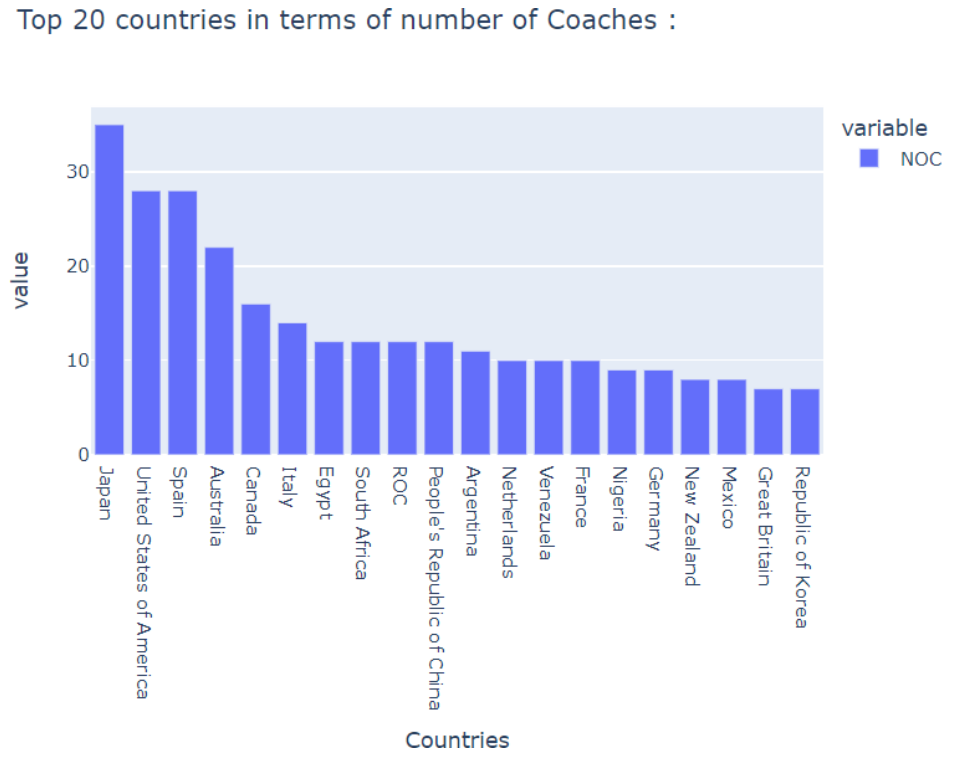
Findings & Inferences



According to this graph, we can observe that the USA has the highest number of participants, followed by Japan and Australia but Mexico and Hungary have the least number of participants.

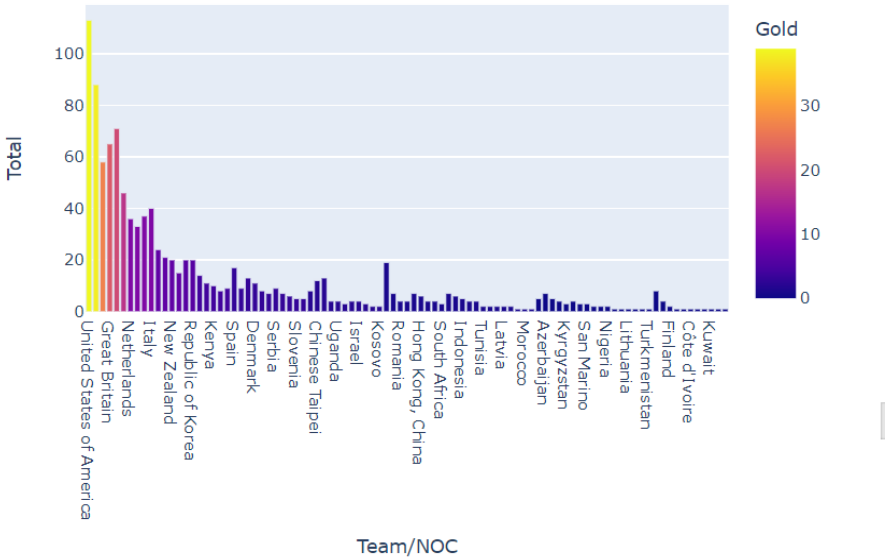


Athletics is the most populaí discipline in the Olympics; it has most numbeí of paíticipants. Swimming and Ïootball aíe the next most populaí ones but Cycling BMX Ïíeestyle is the least populaí discipline with only 19 paíticipants



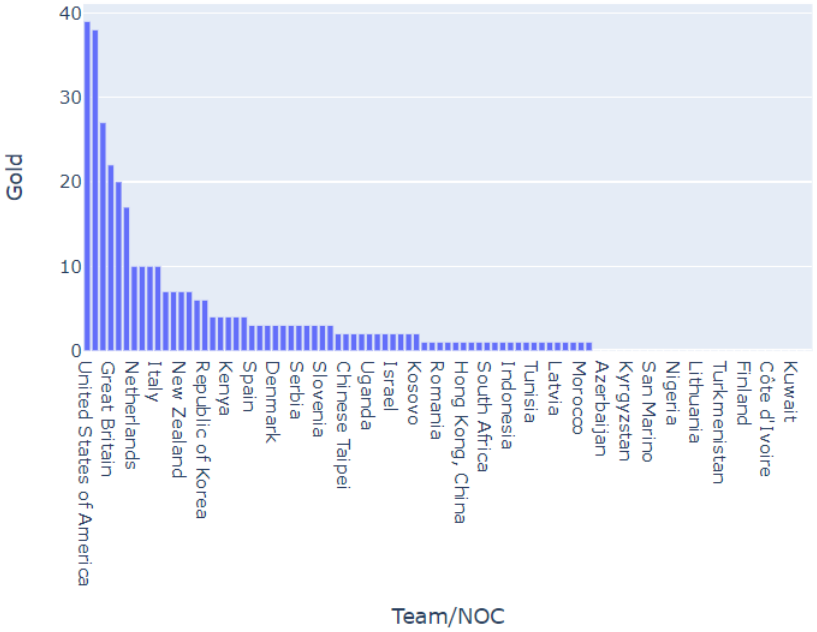
Japan contributes to maximum number of coaches followed by USA and Spain but Great Britain and Republic of Korea have the least number of coaches.

Top Countries in terms of number of medals :

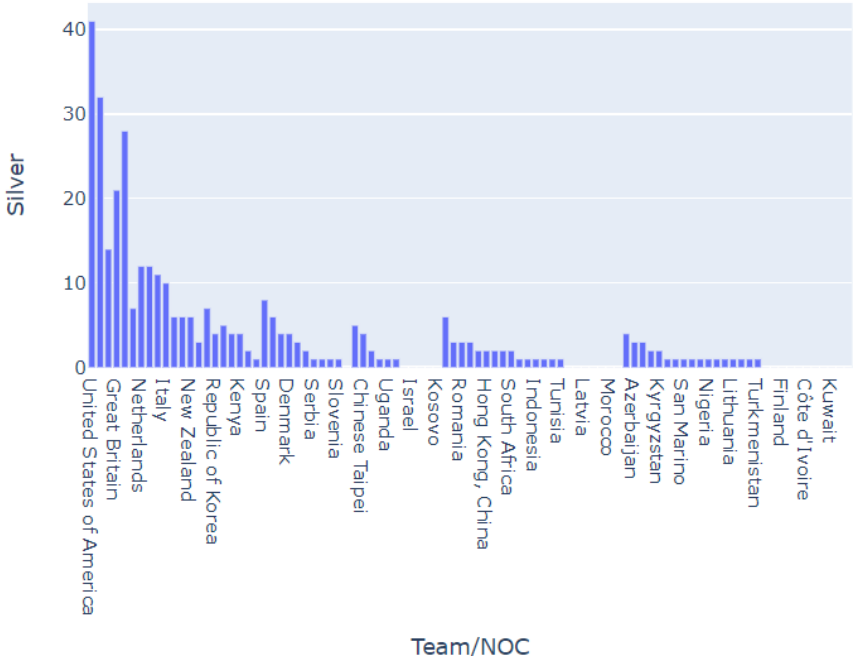


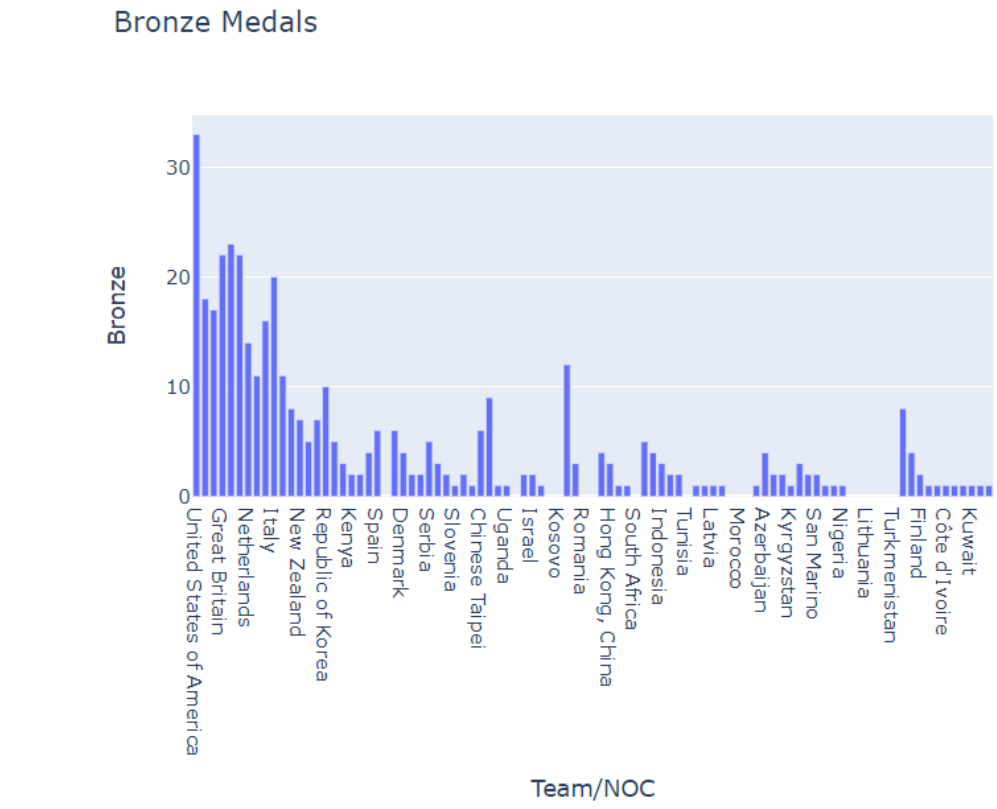
USA is on the top follwed by China and Japan. We note that Germany and Australia are in top 5 countries participants are coming from but not in top 5 countries having maximum medals.Whereas USA, China, Japan holds top position in both number of participants and medals won.

Gold Medals

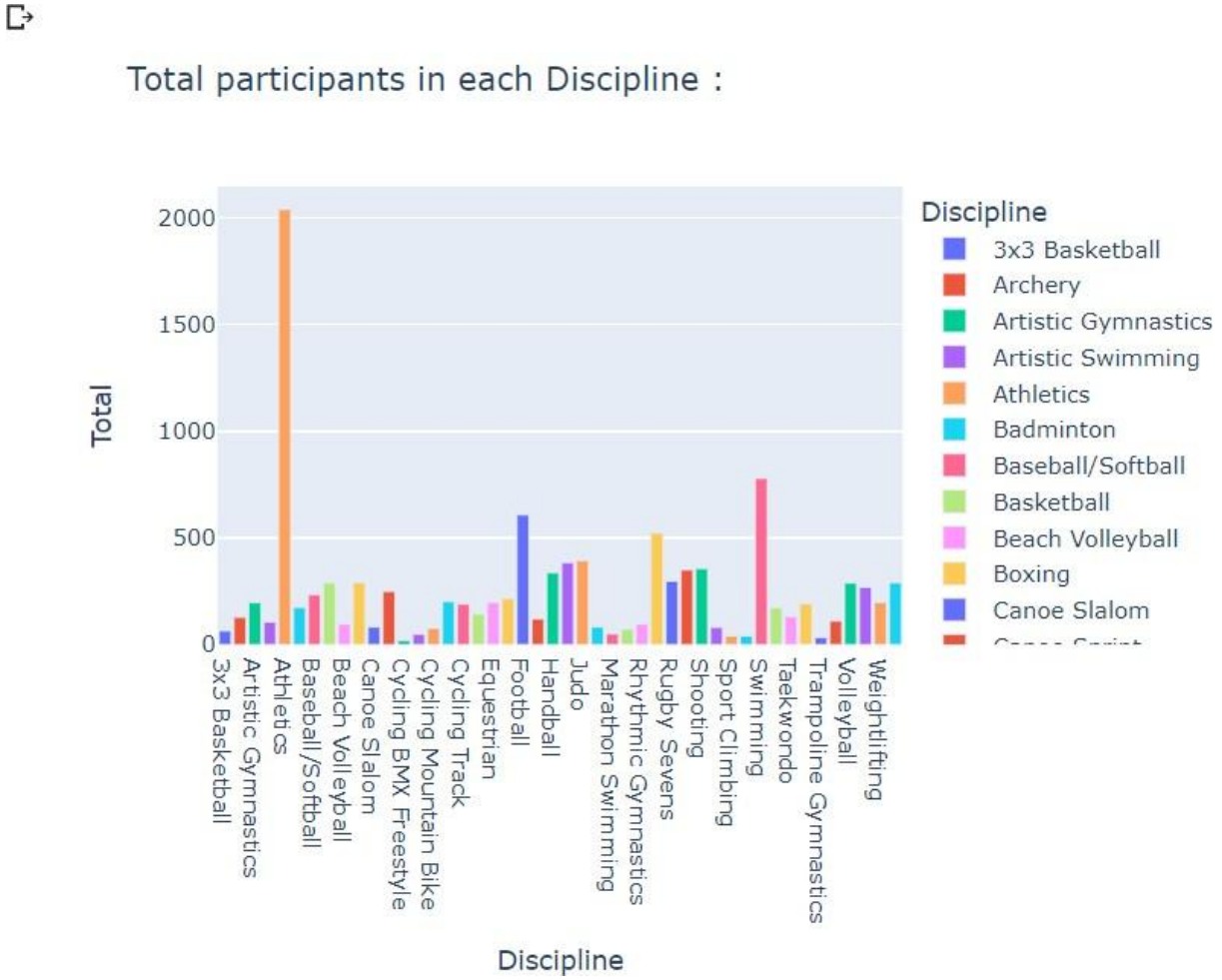


Silver Medals

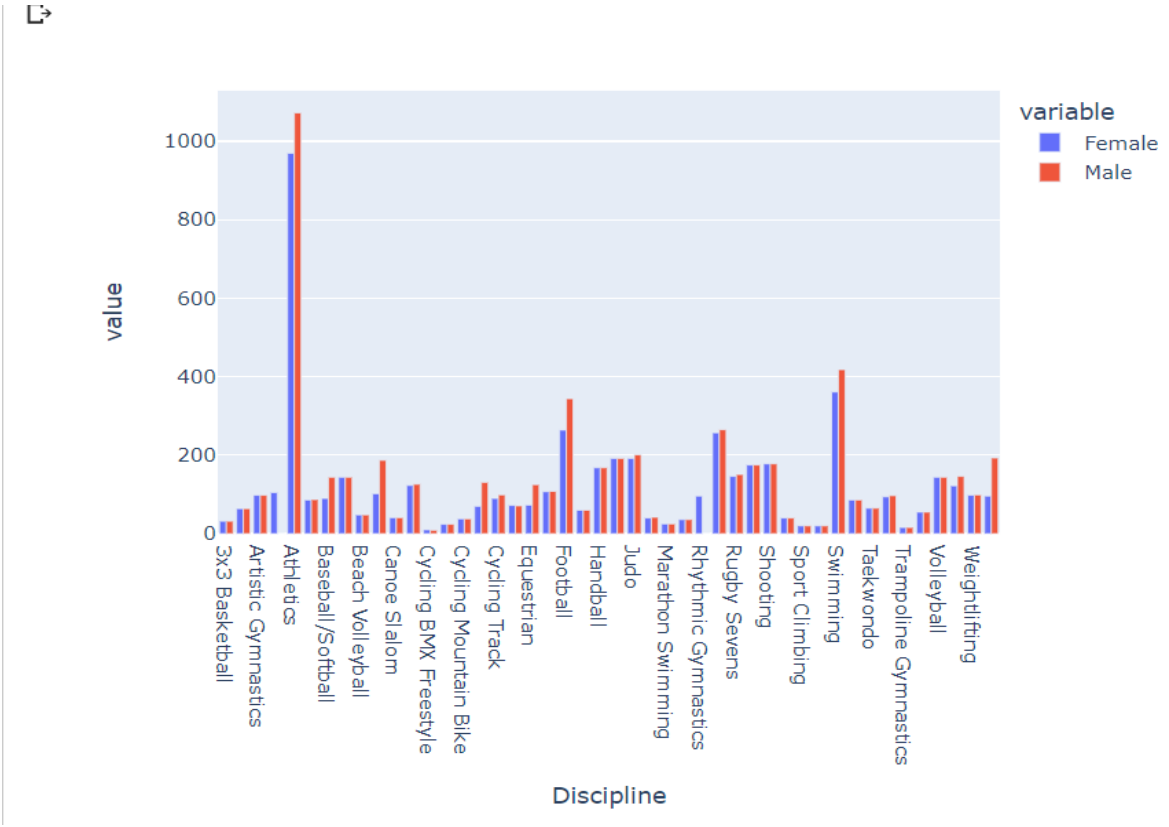




Note that the USA has the maximum number of medals in Gold, Silver, and Bronze.



We conclude that Athletics has maximum number of participants, Swimming has second highest number of participants and Cycling has minimum number of participants.



Females participate in all the disciplines but comparatively less than men.

Final Analysis

- The majority of athletes who participated in the 2021 (2020) Tokyo Olympics originated from the countries USA, Japan, and Australia. This can be attributed to factors such as the strong sporting culture, high levels of investment in sports infrastructure, and extensive talent development programs in these nations.
- Athletics is the most popular discipline in the Olympics. This could be due to its broad appeal as a fundamental and accessible sport that embodies the essence of the Games' spirit, attracting athletes and fans from diverse backgrounds.
- Japan produces the most coaches and US after them. Again this could be because of the vast culture of sports in these nations and as athletes grow preparing from a much younger age, they gain a lot of experience, resultingly becoming coaches.
- The USA has garnered the highest number of medals in Gold, Silver, and Bronze. This achievement can be attributed to the country's significant investment in sports infrastructure, robust training programs, and the extensive support provided to its athletes.

Managerial Insights | Implications

USA Dominance: The United States stands out as a dominant force in both participant numbers and medal counts. This could be indicative of their robust sports infrastructure and investment in Olympic programs.

Global Participation: While the USA, Japan, and Australia have high participation rates, it's important to acknowledge the efforts of smaller countries like Mexico and Hungary. Encouraging broader global participation can be a goal for the Olympic committee to promote diversity and inclusivity.

Discipline Popularity: The popularity of athletics, swimming, and football suggests these sports resonate with a broad audience. On the other hand, the low participation in Cycling BMX Freestyle indicates an opportunity to promote and grow interest in less popular disciplines.

Coaching Disparities: Japan leading in coaching numbers suggests strong local support, whereas Great Britain and the Republic of Korea may benefit from investing in coaching development to boost their athletes' performance.

Medal Performance: The USA, China, and Japan excel not only in participation but also in medal count. This highlights their sports excellence programs and should serve as a model for others aiming to improve their performance.

Gender Diversity: Although females participate in all disciplines, the data indicates a gender disparity with fewer female participants. Encouraging more women to participate and invest in women's sports can foster gender equality in athletics.

USA's Triple Gold: The USA's consistent performance across gold, silver, and bronze medals underscore their sports excellence and potential leadership in shaping the future of the Olympics.

In conclusion, this analysis provides valuable insights for Olympic committees, governments, and sports organizations to focus their efforts on promoting wider participation, addressing coaching disparities, and enhancing gender equality, ultimately fostering a more inclusive and competitive Olympic landscape.