Tokyo olympics 2021

Tanya Jagyasi (202118039) Aastha Deogharia (202118025)

I. Introduction

In this project we were using Tokyo Olympic 2021 database .With this project we try to analyze how the Olympics have evolved over time, including questions about the participation and performance of women/men, the variation in athletic build by sport and by nation, and the nature of olympic dominance by nation, which is demonstrated by number of medals received.

II. Dataset

This dataset is in comma separated format.Which contains 5 tables namely.

- Athlete.csv— Contains details about the participated athletes
- Coaches.csv — Contains details about the coaches (country, discipline, event)
- EntierGender.csv— Entries by Discipline and number of female and male athletes taking part in it
- Medal.csv— Number of gold, silver and bronze medals won by each country
- Team.csv— Contains the details of all the Teams(Country, event, Discipline, Event)

III. OLAP

OLAP (for online analytical processing) is software for performing multidimensional analysis at high speeds on large volumes of data from a data warehousing, data mart, or some other unified, centralized data store. Most business data have multiple dimensions—multiple categories into which the data are broken down for presentation, tracking, or analysis.
For example, Tokyo Olympic 2021 database might have several dimensions related to Team (Name,Discipline,NOC, Event), Medal (Rank,Team/NOC,Gold,Silver,Bronze,Total,RankbyTotal),Gender (Discipline,Female,Male,Total), and more. But in a data warehouse, data sets are stored in tables, each of which can organize data into just two of these dimensions at a time. OLAP extracts data from multiple relational data sets and reorganizes it into a multidimensional format that enables very fast processing and very insightful analysis.

IV. Data PreProcessing

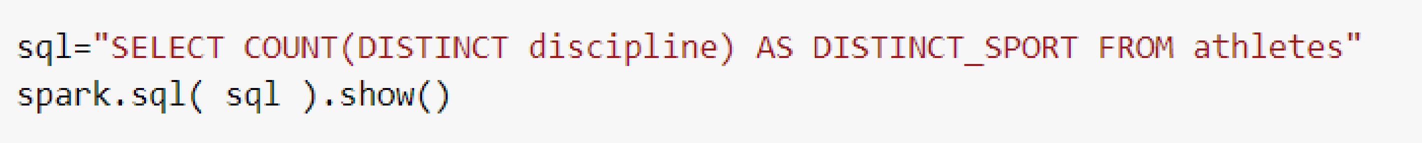
PySpark is the collaboration of Apache Spark and Python.
Apache Spark is an open-source cluster-computing framework, built around speed, ease of use, and streaming analytics whereas Python is a general-purpose, high-level programming language. It provides a wide range of libraries and is majorly used for Machine Learning and Real-Time Streaming Analytics.
In other words, it is a Python API for Spark that lets you harness the simplicity of Python and the power of Apache Spark in order to tame Big Data.
We used PIPy for installation of PySpark.
Spark Session :- SparkSession introduced in version 2.0, It is an entry point to underlying PySpark functionality in order to programmatically create PySpark RDD, DataFrame. Spark session helps im getting Dataframe as tables, execute SQL over tables, cache tables, and read parquet files.
We have created a function to convert CSV to Tables , it contains the name of the csv and name of the table to which we converted the csv.
We have shown 10 elements of all the tables.

V. Analyzing the data using OLAP

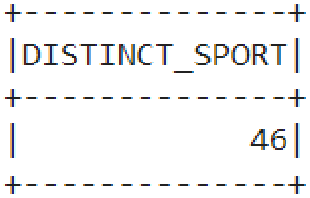
By using OLAP and Aggregation operations we answered the followings questions:

1. How many different sports played in Tokyo Olympic 2021

Answer :– There are total 46 different games played in Tokyo Olympic 2021



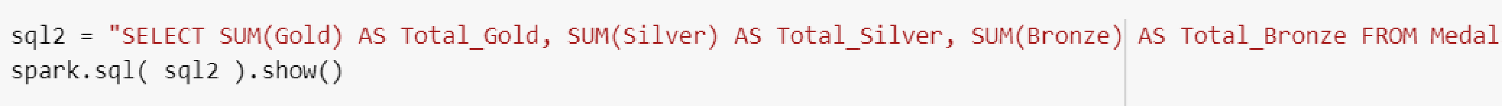
```
sql="SELECT COUNT(DISTINCT discipline) AS DISTINCT_SPORT FROM athletes"
spark.sql( sql ).show()
```



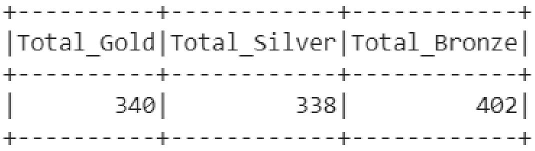
```
+--------------+
|DISTINCT_SPORT|
+--------------+
|            46|
+--------------+
```

2. What is the total number of Gold Silver and Bronze Medals given in Tokyo Olympic 2021

Answer :–There are total 340 Gold Medal ,338 Silver Medal and 402 Bronze Medals were awarded to the winners in Tokyo Olympic 2021

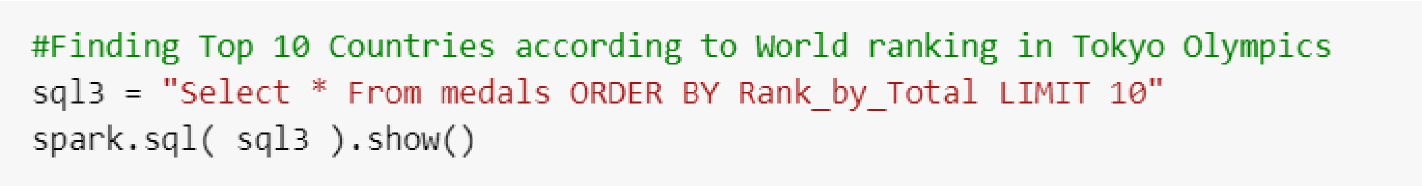


```
sql2 = "SELECT SUM(Gold) AS Total_Gold, SUM(Silver) AS Total_Silver, SUM(Bronze) AS Total_Bronze FROM Medal
spark.sql( sql2 ).show()
```

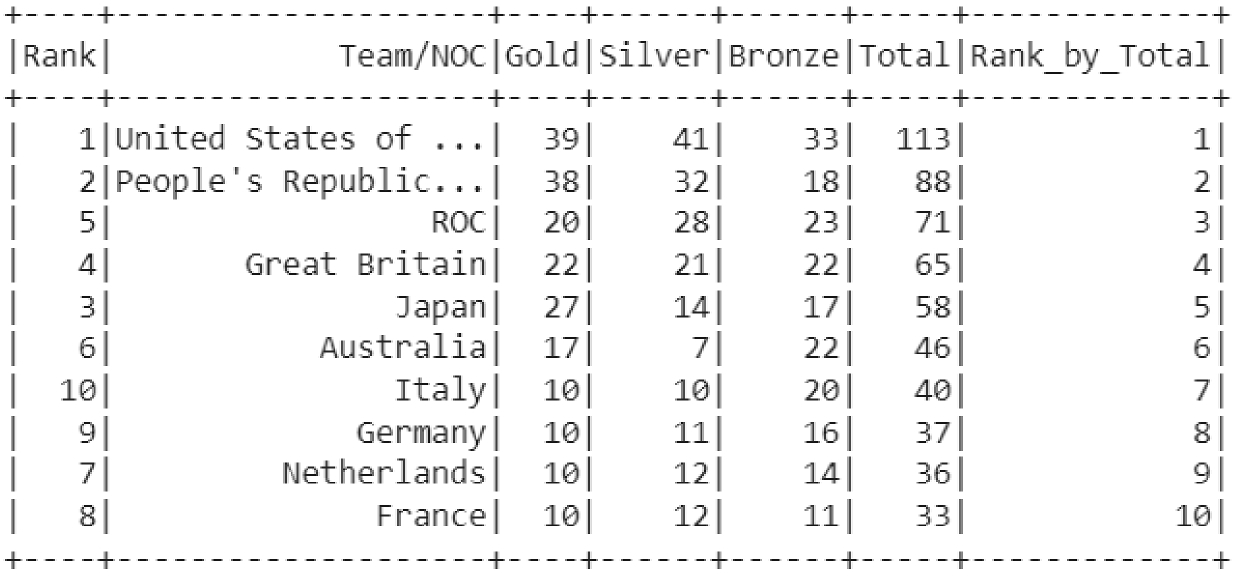


```
+----------+------------+------------+
|Total_Gold|Total_Silver|Total_Bronze|
+----------+------------+------------+
|       340|         338|         402|
+----------+------------+------------+
```

3. List of the top ten Countries with most Medals

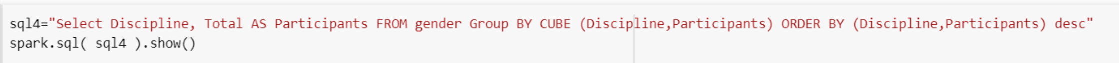


```
#Finding Top 10 Countries according to World ranking in Tokyo Olympics
sql3 = "Select * From medals ORDER BY Rank_by_Total LIMIT 10"
spark.sql( sql3 ).show()
```

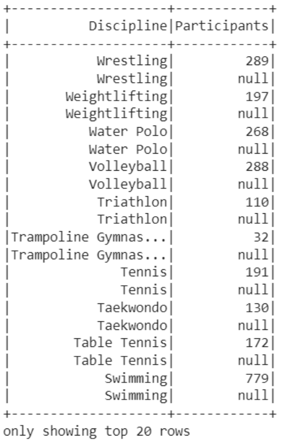


```
+----+--------------------+----+------+------+-----+-------------+
|Rank|            Team/NOC|Gold|Silver|Bronze|Total|Rank_by_Total|
+----+--------------------+----+------+------+-----+-------------+
|   1|United States of ...|  39|    41|    33|  113|            1|
|   2|People's Republic...|  38|    32|    18|   88|            2|
|   5|                 ROC|  20|    28|    23|   71|            3|
|   4|       Great Britain|  22|    21|    22|   65|            4|
|   3|               Japan|  27|    14|    17|   58|            5|
|   6|           Australia|  17|     7|    22|   46|            6|
|  10|               Italy|  10|    10|    20|   40|            7|
|   9|             Germany|  10|    11|    16|   37|            8|
|   7|         Netherlands|  10|    12|    14|   36|            9|
|   8|              France|  10|    12|    11|   33|           10|
+----+--------------------+----+------+------+-----+-------------+
```

4.Numbers of participants in the different Games .

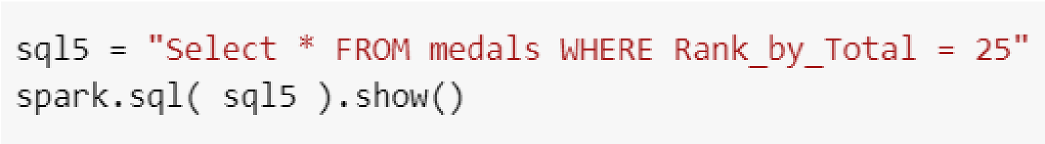


```
sql4="Select Discipline, Total AS Participants FROM gender Group BY CUBE (Discipline,Participants) ORDER BY (Discipline,Participants) desc"
spark.sql( sql4 ).show()
```

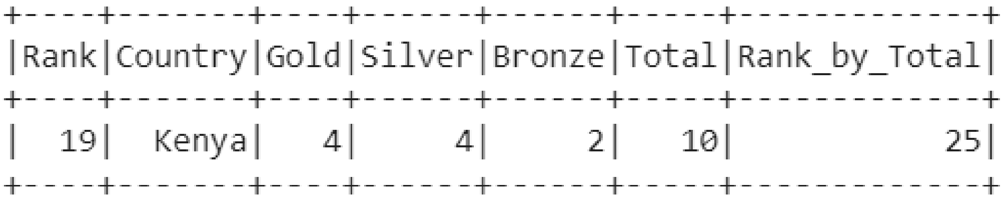


```
+--------------------+------------+
|          Discipline|Participants|
+--------------------+------------+
|           Wrestling|         289|
|           Wrestling|        null|
|       Weightlifting|         197|
|       Weightlifting|        null|
|         Water Polo|         268|
|         Water Polo|        null|
|          Volleyball|         288|
|          Volleyball|        null|
|           Triathlon|         110|
|           Triathlon|        null|
|Trampoline Gymnas...|          32|
|Trampoline Gymnas...|        null|
|              Tennis|         191|
|              Tennis|        null|
|           Taekwondo|         130|
|           Taekwondo|        null|
|        Table Tennis|         172|
|        Table Tennis|        null|
|            Swimming|         779|
|            Swimming|        null|
+--------------------+------------+
only showing top 20 rows
```

5. Which country scored 25th in the World Ranking

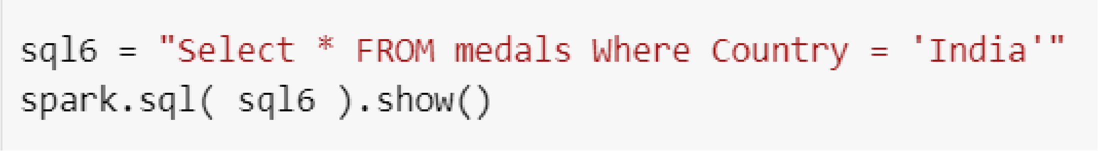


```
sql5 = "Select * FROM medals WHERE Rank_by_Total = 25"
spark.sql( sql5 ).show()
```

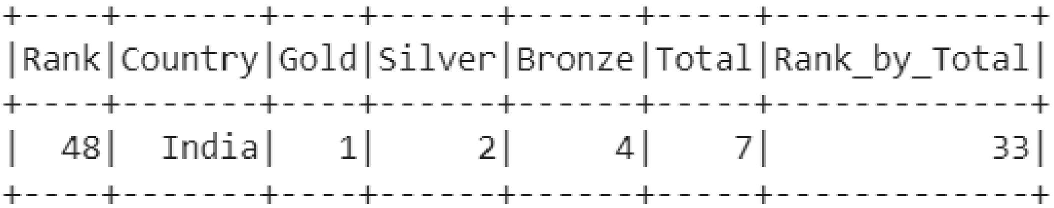


```
+----+-------+----+------+------+-----+-------------+
|Rank|Country|Gold|Silver|Bronze|Total|Rank_by_Total|
+----+-------+----+------+------+-----+-------------+
|  19|  Kenya|   4|     4|     2|   10|           25|
+----+-------+----+------+------+-----+-------------+
```

6. Total number of Medals won by “India” .

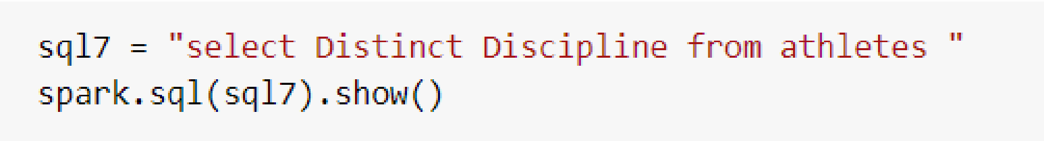


```
sql6 = "Select * FROM medals Where Country = 'India'"
spark.sql( sql6 ).show()
```

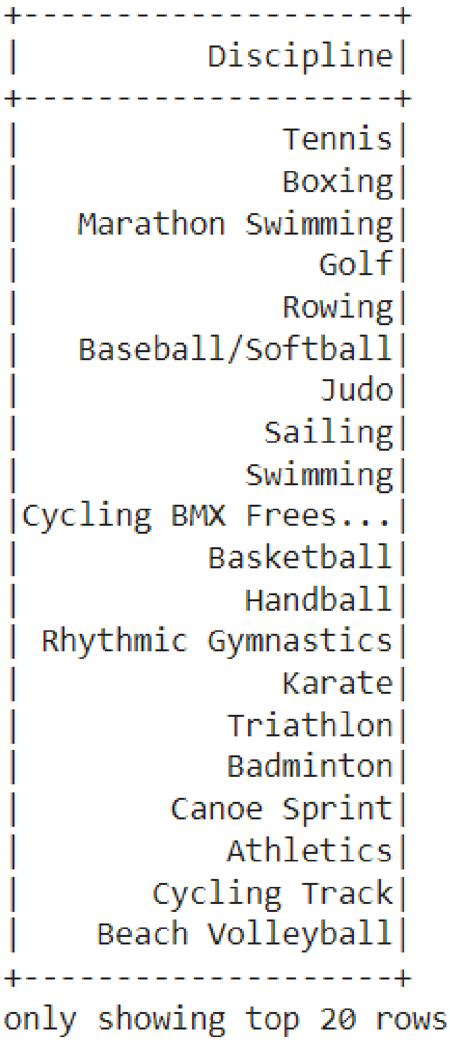


```
+----+-------+----+------+------+-----+-------------+
|Rank|Country|Gold|Silver|Bronze|Total|Rank_by_Total|
+----+-------+----+------+------+-----+-------------+
|  48|  India|   1|     2|     4|    7|           33|
+----+-------+----+------+------+-----+-------------+
```

7. What are the different sports played in Tokyo Olympic 2021

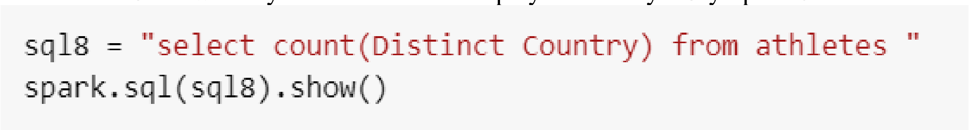


```
sql7 = "select Distinct Discipline from athletes "
spark.sql(sql7).show()
```

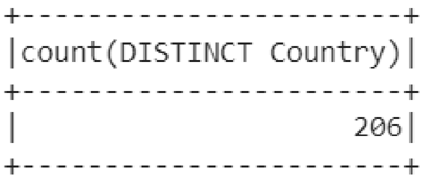


```
+--------------------+
|          Discipline|
+--------------------+
|              Tennis|
|              Boxing|
|   Marathon Swimming|
|                Golf|
|              Rowing|
|   Baseball/Softball|
|                Judo|
|             Sailing|
|            Swimming|
|Cycling BMX Frees...|
|          Basketball|
|            Handball|
| Rhythmic Gymnastics|
|              Karate|
|           Triathlon|
|           Badminton|
|        Canoe Sprint|
|           Athletics|
|       Cycling Track|
|    Beach Volleyball|
+--------------------+
only showing top 20 rows
```

8. How many different countries played in Tokyo Olympic 2021

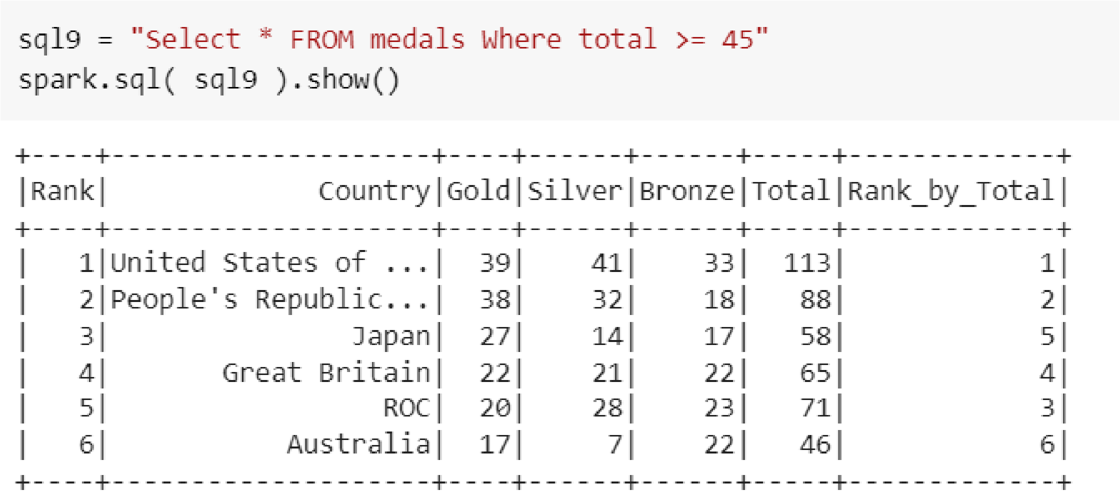


```
sql8 = "select count(Distinct Country) from athletes "
spark.sql(sql8).show()
```



```
+----------------------+
|count(DISTINCT Country)|
+----------------------+
|                   206|
+----------------------+
```

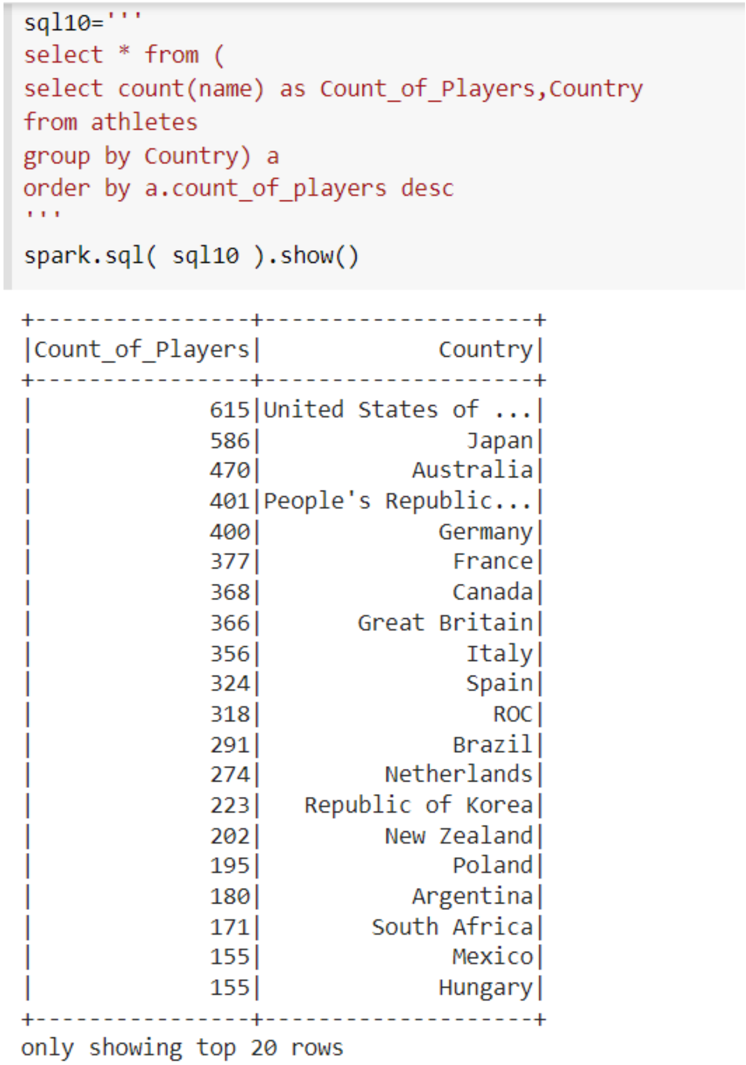
9. Countries who won more than 45 medals in the Tokyo Olympic 2021



```
sql9 = "Select * FROM medals Where total >= 45"
spark.sql( sql9 ).show()
```

```
+----+--------------------+----+------+------+-----+-------------+
|Rank|             Country|Gold|Silver|Bronze|Total|Rank_by_Total|
+----+--------------------+----+------+------+-----+-------------+
|   1|United States of ...|  39|    41|    33|  113|            1|
|   2|People's Republic...|  38|    32|    18|   88|            2|
|   3|               Japan|  27|    14|    17|   58|            5|
|   4|       Great Britain|  22|    21|    22|   65|            4|
|   5|                 ROC|  20|    28|    23|   71|            3|
|   6|           Australia|  17|     7|    22|   46|            6|
+----+--------------------+----+------+------+-----+-------------+
```

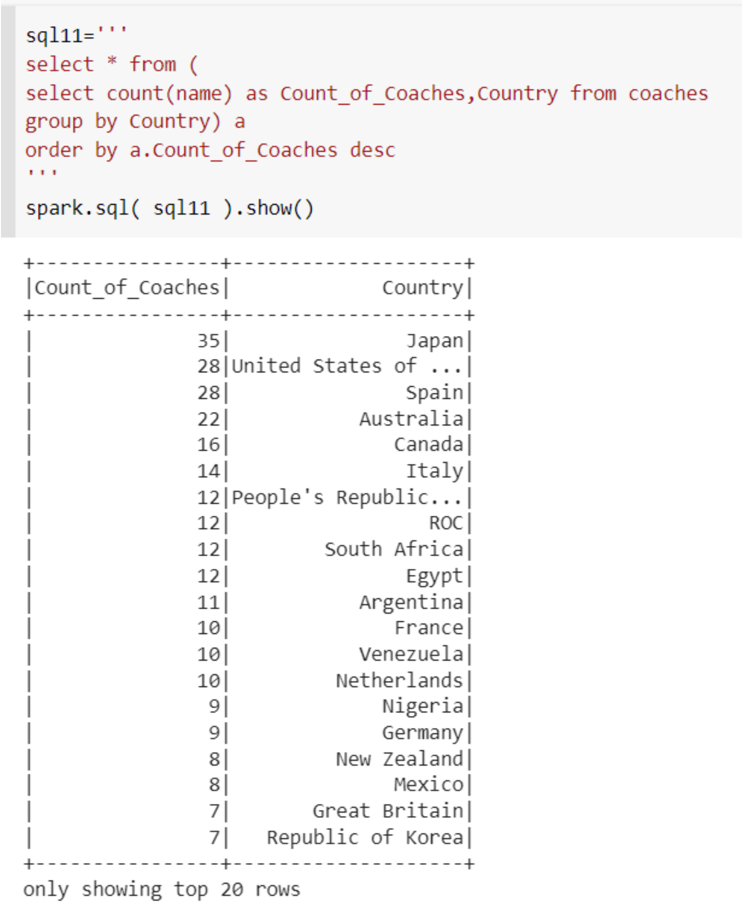
10. Total number of participants across the countries



```
sql10='''
select * from (
select count(name) as Count_of_Players,Country
from athletes
group by Country) a
order by a.count_of_players desc
'''
spark.sql( sql10 ).show()
```

```
+----------------+--------------------+
|Count_of_Players|             Country|
+----------------+--------------------+
|             615|United States of ...|
|             586|               Japan|
|             470|           Australia|
|             401|People's Republic...|
|             400|             Germany|
|             377|              France|
|             368|              Canada|
|             366|       Great Britain|
|             356|               Italy|
|             324|               Spain|
|             318|                 ROC|
|             291|              Brazil|
|             274|         Netherlands|
|             223|   Republic of Korea|
|             202|         New Zealand|
|             195|              Poland|
|             180|           Argentina|
|             171|        South Africa|
|             155|              Mexico|
|             155|             Hungary|
+----------------+--------------------+
only showing top 20 rows
```

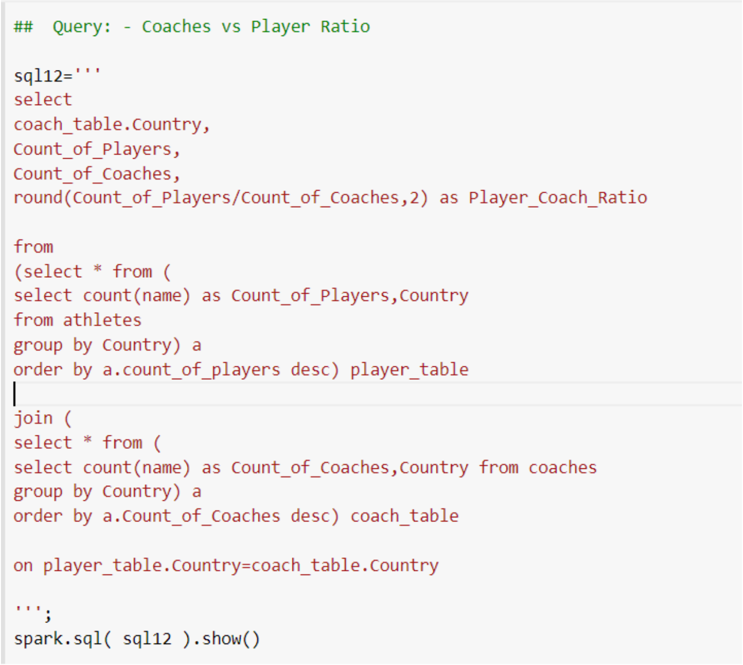
11. Number of coaches from different countries



```
sql11='''
select * from (
select count(name) as Count_of_Coaches,Country from coaches
group by Country) a
order by a.Count_of_Coaches desc
'''
spark.sql( sql11 ).show()
```

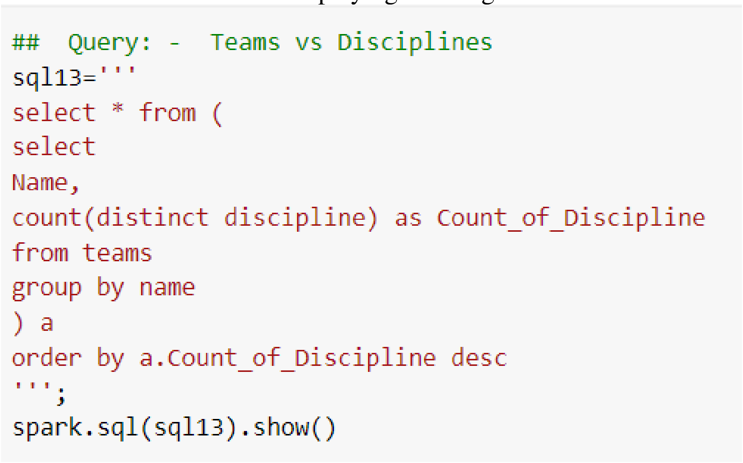
```
+----------------+--------------------+
|Count_of_Coaches|             Country|
+----------------+--------------------+
|              35|               Japan|
|              28|United States of ...|
|              28|               Spain|
|              22|           Australia|
|              16|              Canada|
|              14|               Italy|
|              12|People's Republic...|
|              12|                 ROC|
|              12|        South Africa|
|              12|               Egypt|
|              11|           Argentina|
|              10|              France|
|              10|           Venezuela|
|              10|         Netherlands|
|               9|             Nigeria|
|               9|             Germany|
|               8|         New Zealand|
|               8|              Mexico|
|               7|       Great Britain|
|               7|   Republic of Korea|
+----------------+--------------------+
only showing top 20 rows
```

12. How many coaches were assigned to which number of players.



```
##  Query: - Coaches vs Player Ratio

sql12='''
select
coach_table.Country,
Count_of_Players,
Count_of_Coaches,
round(Count_of_Players/Count_of_Coaches,2) as Player_Coach_Ratio

from
(select * from (
select count(name) as Count_of_Players,Country
from athletes
group by Country) a
order by a.count_of_players desc) player_table

join (
select * from (
select count(name) as Count_of_Coaches,Country from coaches
group by Country) a
order by a.Count_of_Coaches desc) coach_table

on player_table.Country=coach_table.Country

''';
spark.sql( sql12 ).show()
```

13. Teams playing which game.



```
##  Query: -  Teams vs Disciplines
sql13='''
select * from (
select
Name,
count(distinct discipline) as Count_of_Discipline
from teams
group by name
) a
order by a.Count_of_Discipline desc
''';
spark.sql(sql13).show()
```

VI. Conclusion

- Most athletes in the Olympics reside from countries like the United States of America, Japan and Australia. This could be because these countries promote sports and athletics from a younger age, hone the youth and prepare them to compete in such higher levels and invest more behind these. Unlike countries like Tanzania or Sudan, these countries probably have a better sports culture. Athletics is the most popular discipline in the Olympics. This could be because other sports like football/basketball get other stages for players to display their merits so probably this is why athletes performing in athletics try to display their abilities in this grand stage.
- Japan produces the most coaches and the 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, resultantly becoming coaches.
- Females participate in all the disciplines but comparatively less than men. This cannot be because of any kinds of discrimination, rather the lack of growth of mindsets in young women to become great athletes someday in comparison to men. But, surely this statistic has developed a lot in the past years and some day, women might dominate more in these stages too.
- The USA has the highest number of gold, silver and bronze medals and this shows how successful the US is. The biggest reason could be wealth; more investment on athletes and their respective disciplines. Having better coaches and facilities does help create better and improved athletes so wealth could probably be why the US does so much better than other countries