

SQL Data Analysis PROJECT



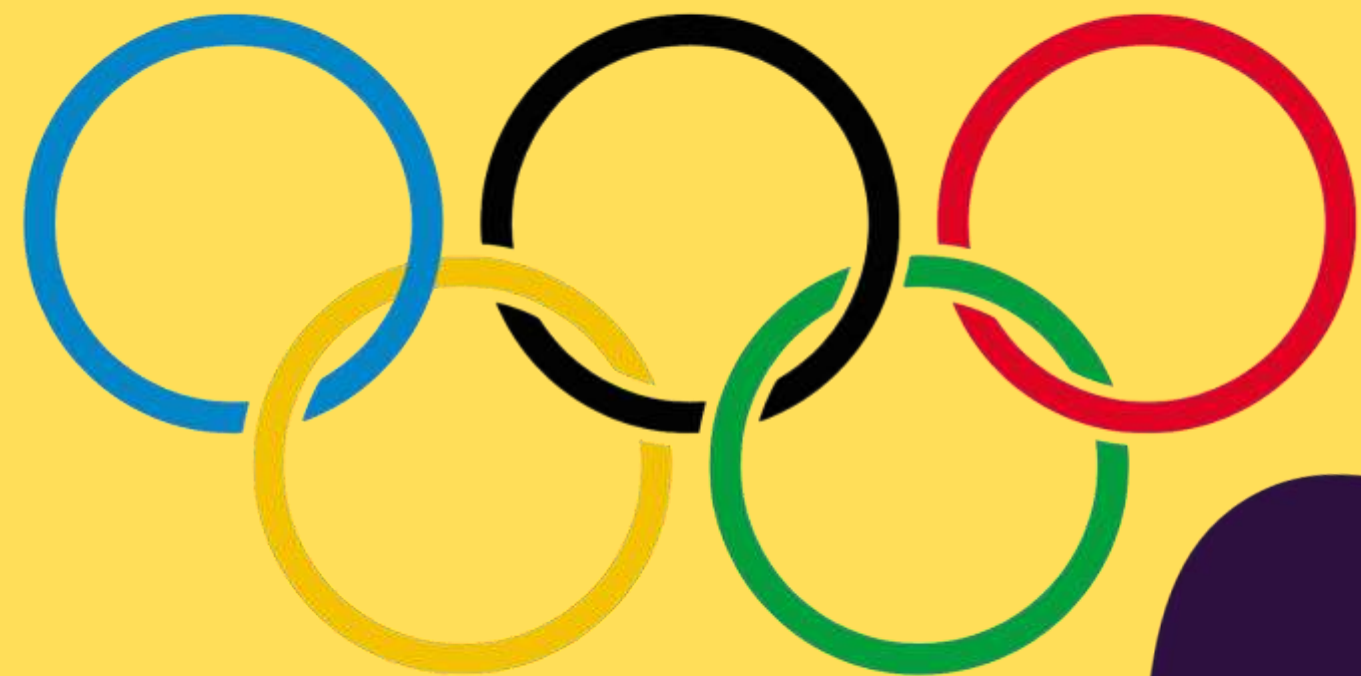
Created by Priyesh



Overview

The project is focused on examining Olympic data using MySQL. We will explore different areas like trends in participation, the performance of countries and athletes, demographic information, analysis of sports and distribution of medals. We will use MySQL queries to present our findings in an organized and easy-to-understand way.

Olympic Data Analysis



Dataset

This is a historical dataset on the modern Olympic Games, including all the Games from Athens 1896 to Rio 2016. We Have two datasets. One Named **olympic_event** and other one is **noc_regions**.

cross-referencing the two tables.

The file **Olympic_events.csv** contains 271116 rows and 15 columns. Each row corresponds to an individual athlete competing in an individual Olympic event.

And the file **noc_regions.csv** contains the **noc** code present in Olympic event and the corresponding region to each **noc**.

We will join the two tables using **noc** whenever required.



Dataset

olympic_event

	name	sex	Age	Height	Weight	Team	Noc	Games	Year	Season	City	Sport	Event	Medal
▶	A Dijiang	M	24	180	80	China	CHN	1992 Summer	1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NA
	A Lamusi	M	23	170	60	China	CHN	2012 Summer	2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NA
	Gunnar Nielsen Aaby	M	24	NA	NA	Denmark	DEN	1920 Summer	1920	Summer	Antwerpen	Football	Football Men's Football	NA
	Edgar Lindenau Aabye	M	34	NA	NA	Denmark/Sweden	DEN	1900 Summer	1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
	Christine Jacoba Aaftink	F	21	185	82	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 m...	NA
	Christine Jacoba Aaftink	F	21	185	82	Netherlands	NED	1988 Winter	1988	Winter	Calgary	Speed Skating	Speed Skating Women's 1,000 ...	NA
	Christine Jacoba Aaftink	F	25	185	82	Netherlands	NED	1992 Winter	1992	Winter	Albertville	Speed Skating	Speed Skating Women's 500 m...	NA
	Christine Jacoba Aaftink	F	25	185	82	Netherlands	NED	1992 Winter	1992	Winter	Albertville	Speed Skating	Speed Skating Women's 1,000 ...	NA
	Christine Jacoba Aaftink	F	27	185	82	Netherlands	NED	1994 Winter	1994	Winter	Lillehammer	Speed Skating	Speed Skating Women's 500 m...	NA

Noc_regions

	NOC	region	notes
	AFG	Afghanistan	
	AHO	Curacao	Netherlands Antilles
	ALB	Albania	
	ALG	Algeria	
	AND	Andorra	
	ANG	Angola	
	ANT	Antigua	Antigua and Barbuda
	ANZ	Australia	Australasia

Importing Datasets

Creating Table Schema:

```
create table Olympic_event(  
  name varchar(1000),  
  sex varchar(1000),  
  Age varchar(1000),  
  Height varchar(1000),  
  Weight varchar(1000),  
  Team varchar(1000),  
  Noc varchar(1000),  
  Games varchar(1000),  
  Year varchar(1000),  
  Season varchar(1000),  
  City varchar(1000),  
  Sport varchar(1000),  
  Event varchar(1000),  
  Medal varchar(1000)  
);
```

Importing Data Into Table:

```
LOAD DATA INFILE 'D:\olympic_event.csv'  
INTO TABLE Olympic_event  
FIELDS TERMINATED BY ','  
ENCLOSED BY '"'  
LINES TERMINATED BY '\n'  
IGNORE 1 ROWS;
```



Problem Statements

1. How many Olympic games have been held?
2. List down all Olympic games held so far along with year, season and city.
3. Mention the total no of nations who participated in each Olympic game?
4. Which year saw the highest and lowest no of countries participating in Olympic
5. Which nation has participated in all of the olympic games
6. Identify the sport which was played in all summer Olympic.
7. Which Sports were just played only once in the Olympic.
8. Fetch the total no of sports played in each olympic games.
9. Fetch oldest athletes to win a gold medal.
10. Find the Ratio of male and female athletes participated in all olympic games.
11. Fetch the top 5 athletes who have won the most gold medals.
12. Fetch the top 5 most successful countries in Olympic. Success is defined by no of medals won.
13. List down total gold, silver and bronze medals won by each country.
14. Identify which country won the most gold, most silver, most bronze medals and the most medals in the same olympic game..
15. Which countries have never won gold medal but have won silver/bronze medals?



Problem Statement - 1

Query

```
-- 1. How many olympics games have been held?
```

```
Select count(distinct games) as Total_Olympics_Held
from olympic_event;
```

Output

	Total_Olympics_Held
▶	51

Approach:

- We use the COUNT function to count the total number of distinct values in the 'games' column.
- The DISTINCT keyword ensures that each game is counted only once, eliminating duplicates.
- We alias the result as 'Total_Olympics_Held' for clarity.

Problem Statement - 2

Query

```
-- 2. List down all Olympics games held so far along with year,season and city.

Select year,season,city
from olympic_event
Group by year,season,city order by year;
```

Output

	year	season	city
▶	1896	Summer	Athina
	1900	Summer	Paris
	1904	Summer	St. Louis
	1906	Summer	Athina
	1908	Summer	London
	1912	Summer	Stockholm
	1920	Summer	Antwerpen
	1924	Summer	Paris
	1924	Winter	Chamonix
	1928	Summer	Amsterdam

Approach:

- We select the 'year', 'season', and 'city' columns from the 'olympic_event' table.
- Using the GROUP BY clause, we group the results by 'year', 'season', and 'city' to aggregate data.
- The ORDER BY clause sorts the results by 'year' in ascending order.

Problem Statement - 3

Query

```
-- 3. Mention the total no of nations who participated in each olympics game?

select games, count(distinct NOC) as total_countries
from olympic_event
group by games;
```

Output

	games	total_countries
▶	1896 Summer	12
	1900 Summer	31
	1904 Summer	15
	1906 Summer	21
	1908 Summer	22
	1912 Summer	29
	1920 Summer	29
	1924 Summer	45
	1924 Winter	19

Approach:

- We select the 'games' column and count the distinct number of National Olympic Committees (NOCs) represented in each game.
- Using the GROUP BY clause, we group the results by 'games' to aggregate data.

Problem Statement - 4

Query

```
-- 4. Which year saw the highest and lowest no of countries participating in olympics

with all_countries as(
    select games,region
    from olympic_event as ae
    join noc_regions as nr
    on ae.noc=nr.noc
    group by 1,2
),
tot_countries as(
    select games,count(region) as total_countries
    from all_countries
    group by games
)

select distinct concat(first_value(games) over(order by total_countries),'-',
    first_value(total_countries) over(order by total_countries)) as lowest_countries,
    concat(first_value(games) over(order by total_countries desc),'-',
    first_value(total_countries) over(order by total_countries desc)) as highest_countries
from tot_countries;
```

Output

	lowest_countries	highest_countries
▶	1896 Summer-12	2016 Summer-204

Approach:

- We start by setting up a Common Table Expression (CTE) called 'all_countries' to connect the 'olympic_event' table with the 'noc_regions' table, linking NOCs to their regions.
- Next, we establish another CTE named 'tot_countries' to calculate the total count of countries participating in each Olympic event.
- Lastly, we retrieve unique game-country pairs arranged in ascending and descending order based on the total countries represented.

Problem Statement - 5

Query

```
-- 5. Which nation has participated in all of the olympic games
with cte as (
select region,count(distinct games) as total_participated_games
  from olympic_event as ae
 join noc_regions as nr
 on ae.noc=nr.noc
 group by region)

select region,total_participated_games
  from cte
 where total_participated_games=(
select count(distinct games)
  from olympic_event);
```

Output

region	total_participated_games
France	51
Italy	51
Switzerland	51
UK	51

Approach:

- We create a Common Table Expression (CTE) named 'cte' to count the total number of Olympic games each region has participated in.
- The CTE joins the 'olympic_event' table with the 'noc_regions' table to map NOCs to their respective regions.
- We then select regions that have participated in the maximum number of distinct Olympic games.

Problem Statement - 6

Query

```
-- 6. Identify the sport which was played in all summer olympics.

select sport, count(Distinct Games) as total_games
from olympic_event
where season like "Summer"
group by 1
having count(Distinct Games)=(select count(distinct games) from olympic_event
                               where season like "Summer");
```

Output

sport	total_games
Athletics	29
Cycling	29
Fencing	29
Gymnastics	29
Swimming	29

Approach:

- We select the 'sport' column and count the distinct number of Olympic events (Games) for each sport.
- Data is filtered to include only records from Summer seasons.
- Results are grouped by the 'sport' column.
- We use the HAVING clause to filter results, retaining only sports played in all Summer Olympic Games.

Problem Statement - 7

Query

```
-- 7. Which Sports were just played only once in the olympics.

select sport, count(Distinct Games) as total_games
from olympic_event
where season like "Summer"
group by 1
having count(Distinct Games)=1;
```

Output

sport	total_games
Aeronautics	1
Basque Pelota	1
Cricket	1
Croquet	1
Ice Hockey	1
Jeu De Paume	1
Motorboating	1
Racquets	1
Roque	1
Rugby Sevens	1

Approach:

- We select the 'sport' column and count the distinct number of games in which each sport has been featured, focusing only on Summer Olympics.
- Using the GROUP BY clause, we group the results by sport.
- The HAVING clause filters the results to include only sports that have been featured in exactly one Summer Olympic Games.

Problem Statement - 8

Query

```
-- 8. Fetch the total no of sports played in each olympic games.  
  
Select games,count(distinct Sport) as total_sports_Played  
from olympic_event  
group by 1;
```

Output

games	total_sports_Played
1896 Summer	9
1900 Summer	20
1904 Summer	18
1906 Summer	13
1908 Summer	24
1912 Summer	17
1920 Summer	25

Approach:

- We select the 'games' column and count the distinct number of sports played in each Olympic game.
- Using the GROUP BY clause, we group the results by the 'games' column.

Problem Statement - 9

Query

```
-- 9.Fetch oldest athletes to win a gold medal

with temp as(select name,sex,cast(case when age='NA' then '0' else age end as unsigned) as age ,
               team,games,city,sport, event, medal
               from olympic_event),

ranking as(select *,rank() over( order by age desc) as rnk
            from temp where medal like '%Gold%')

select * from ranking
where rnk=1;
```

Output

	name	sex	age	team	games	city	sport	event	medal	rnk
▶	Oscar Gomer Swahn	M	64	Sweden	1912 Summer	Stockholm	Shooting	Shooting Men's Running Target...	Gold	1
	Charles Jacobus	M	64	United States	1904 Summer	St. Louis	Roque	Roque Men's Singles	Gold	1

Approach:

- We create a temporary table 'temp' to handle the data transformation, casting the 'age' column to an unsigned integer and replacing 'NA' values with '0'.
- Then, we create another temporary table 'ranking' to rank the athletes by age in descending order, filtering for gold medalists only.
- Finally, we select the top-ranked athlete with the oldest age.

Problem Statement – 10

Query

```
-- 10. Find the Ratio of male and female athletes participated in all olympic games.
with total_count as
    (select sex, count(1) as cnt
     from olympic_event
     group by sex),
ranking as
    (select *, row_number() over(order by cnt) as rn
     from total_count),
min_cnt as
    (select cnt from ranking where rn = 1),
max_cnt as
    (select cnt from ranking where rn = 2)
select concat('1 : ', round(max_cnt.cnt/min_cnt.cnt, 2)) as ratio
from min_cnt, max_cnt;
```

Output

	ratio
▶	1 : 2.64

Approach:

- We first create a Common Table Expression (CTE) named 'total_count' to count the number of participants by gender.
- Then, we create another CTE named 'ranking' to assign row numbers based on the participant count.
- Next, we create two CTEs, 'min_cnt' and 'max_cnt', to extract the participant counts for the genders with the lowest and highest counts respectively.
- Finally, we calculate the ratio of male to female participation and display it as a formatted string.

Problem Statement - 11

Query

```
-- 11. Fetch the top 5 athletes who have won the most gold medals.
with temp as(
select name,team,(case when
                    medal like "%Gold%" then 1 else 0 end )as gold_count from olympic_event)

select name,sum(gold_count) as total_gold_medals
from temp
group by name
order by sum(gold_count) desc
limit 5;
```

Output

name	total_gold_medals
Michael Fred Phelps, II	23
Raymond Clarence "Ray" Ewry	10
Frederick Carlton "Carl" Lewis	9
Larysa Semenivna Latynina (Diriy-)	9
Paavo Johannes Nurmi	9

Approach:

- We create a temporary table 'temp' to handle the data transformation, where we assign a value of 1 to 'gold_count' if the athlete won a gold medal, and 0 otherwise.
- Then, we select the top 5 athletes based on the sum of their gold medal counts, grouping by athlete name and ordering the results in descending order.
- We limit the results to the top 5 athletes.

Problem Statement - 12

Query

```
-- 12. Fetch the top 5 most successful countries in olympics. Success is defined by no of medals won.
with temp as(
    select nr.region,case
        when medal like '%gold%' then 1
        when medal like '%silver%' then 1
        when medal like '%bronze%' then 1
        else 0
    end as medal_count
    from olympic_event as oe
    join noc_regions as nr
    on oe.noc=nr.noc)

select region,sum(medal_count) as total_medals
from temp
group by region
order by total_medals desc
limit 5;
```

Output

region	total_medals
USA	5637
Russia	3947
Germany	3756
UK	2068
France	1777

Approach:

- We create a temporary table 'temp' to handle the data transformation, where we assign a value of 1 to 'medal_count' if the region won any medal (gold, silver, or bronze), and 0 otherwise.
- Then, we select the top 5 regions based on the sum of their medal counts, grouping by region and ordering the results in descending order.
- We limit the results to the top 5 regions.

Problem Statement - 13

Query

```
-- 13. List down total gold, silver and bronze medals won by each country.

with temp as(
  select games,nr.region,case when medal like '%gold%' then 1 else 0 end as gold_count,
         case when medal like '%silver%' then 1 else 0 end as silver_count,
         case when medal like '%bronze%' then 1 else 0 end as bronze_count
  from olympic_event as oe
  join noc_regions as nr
  on oe.noc=nr.noc)

select games,region,sum(gold_count) as gold,sum(silver_count) as silver,sum(bronze_count) as bronze
from temp
group by games,region
order by 1,2;
```

Output

games	region	gold	silver	bronze
1896 Summer	Australia	2	0	1
1896 Summer	Austria	2	1	2
1896 Summer	Denmark	1	2	3
1896 Summer	France	5	4	2
1896 Summer	Germany	25	5	2
1896 Summer	Greece	10	18	20
1896 Summer	Hungary	2	1	3

Approach:

- We create a temporary table 'temp' to handle the data transformation, where we assign values of 1 to 'gold_count', 'silver_count', and 'bronze_count' based on the type of medal won by each region in each game.
- Then, we select the total count of gold, silver, and bronze medals won by each region in each game, grouping by games and region, and ordering the results.

Problem Statement – 14

Query

```
-- 14. Identify which country won the most gold, most silver, most bronze medals and the most medals in the same olympic game.
with temp as(
select games,nr.region,case when medal like '%gold%' then 1 else 0 end as gold_count,
      case when medal like '%silver%' then 1 else 0 end as silver_count,
      case when medal like '%bronze%' then 1 else 0 end as bronze_count,
      case when medal like '%bronze%' or '%silver%' or '%gold%' then 1 else 0 end as total_medal_count
from olympic_event as oe
join noc_regions as nr
on oe.noc=nr.noc),

total_medals as(
select games,region,sum(gold_count) as gold,sum(silver_count) as silver,sum(bronze_count) as bronze,sum(total_medal_count) as total_medal
from temp
group by 1,2),

final_medals as(
select games,row_number() over(partition by games order by gold desc) as gold_rn,concat(region,'-',gold) as max_gold,
row_number() over(partition by games order by silver desc) as silver_rn,concat(region,'-',silver) as max_silver,
row_number() over(partition by games order by bronze desc) as bronze_rn,concat(region,'-',bronze) as max_bronze,
row_number() over(partition by games order by total_medal desc) as max_rn,concat(region,'-',total_medal) as max_medal
from total_medals)

select distinct games,max_gold,max_silver,max_bronze,max_medal
from final_medals
where gold_rn=1 and silver_rn=1 and bronze_rn=1 and max_rn=1
order by games;
```

Output

games	max_gold	max_silver	max_bronze	max_medal
1904 Summer	USA-128	USA-141	USA-125	USA-125
1906 Summer	Greece-24	Greece-48	Greece-30	Greece-30
1908 Summer	UK-147	UK-131	UK-90	UK-90
1932 Summer	USA-81	USA-47	USA-61	USA-61
1936 Summer	Germany-93	Germany-70	Germany-61	Germany-61
1956 Summer	Russia-68	Russia-46	Russia-55	Russia-55
1980 Summer	Russia-187	Russia-129	Russia-126	Russia-126
2008 Summer	USA-127	USA-110	USA-80	USA-80

Approach:

- We first create a Common Table Expression (CTE) named 'temp' to calculate the medal counts for each region in each Olympic game.
- Then, we create a CTE named 'total_medals' to aggregate the medal counts by region for each game.
- Next, we create another CTE named 'final_medals' to assign row numbers for each region based on their medal counts.
- Finally, we select the leading region in terms of gold, silver, bronze, and total medals for each game.

Problem Statement - 15

Query

```
-- 15. Which countries have never won gold medal but have won silver/bronze medals?

with temp as(
select nr.region,case when medal like '%gold%' then 1 else 0 end as gold_count,
       case when medal like '%silver%' then 1 else 0 end as silver_count,
       case when medal like '%bronze%' then 1 else 0 end as bronze_count
from olympic_event as oe
join noc_regions as nr
on oe.noc=nr.noc),

total_medals as(select
region,sum(gold_count) as gold,sum(silver_count) as silver,sum(bronze_count) as bronze
from temp
group by region
order by 1)

select region,gold,silver,bronze
from total_medals
where gold=0 and (silver>0 or bronze>0)
order by region;
```

Output

region	gold	silver	bronze
Afghanistan	0	0	2
Barbados	0	0	1
Bermuda	0	0	1
Botswana	0	1	0
Curacao	0	1	0
Cyprus	0	1	0
Djibouti	0	0	1

Approach

- We create a temporary table 'temp' to handle the data transformation, where we assign values of 1 to 'gold_count', 'silver_count', and 'bronze_count' based on the type of medal won by each region in Olympic events.
- Then, we create another temporary table 'total_medals' to calculate the total count of gold, silver, and bronze medals won by each region, grouping by region.
- Finally, we select regions that have not won any gold medals but have won either silver or bronze medals, ordering the results by region.



Key Findings



- **Total Number of Olympic Games:** The analysis revealed the total number of Olympic Games held so far.
- **List of Olympic Games:** A comprehensive list of all Olympic Games held to date was compiled, including details such as the year, season (Summer or Winter), and host city.
- **Participating Nations:** The total number of nations that participated in each Olympic Games was determined, shedding light on the global representation in the events.
- **Highest and Lowest Participation:** The analysis identified the years with the highest and lowest number of countries participating in the Olympic Games, providing insights into the variability of participation over time.
- **Consistent Participant Nation:** A nation that has participated in all Olympic Games was identified, showcasing its enduring commitment to the event.
- **Universal Summer Olympic Sport:** A sport that has been played in all Summer Olympic Games was identified, highlighting its consistent presence throughout the history of the event.
- **One-Time Olympic Sports:** Sports that were played only once in the Olympic Games were identified, revealing unique or lesser-known events that have been featured in the past.
- **Total Sports Played:** The total number of sports played in each Olympic Games was determined, showcasing the diversity of athletic disciplines featured in the events.



Key Findings



- **Oldest Gold Medalists:** The oldest athletes to win gold medals in the Olympic Games were identified, showcasing remarkable achievements at an advanced age.
- **Gender Ratio:** The ratio of male and female athletes participating in all Olympic Games was calculated, providing insights into gender representation in the events.
- **Top Gold Medalists:** The top five athletes who have won the most gold medals in Olympic history were identified, highlighting their exceptional performances.
- **Most Successful Countries:** The top five most successful countries in terms of total medals won were determined, showcasing their dominance in the Olympic Games.
- **Medal Distribution by Country:** The total number of gold, silver, and bronze medals won by each country was compiled, providing a comprehensive overview of medal distribution.
- **Country Achievement in Each Game:** The countries that won the most gold, silver, bronze, and overall medals in each Olympic Games were identified, showcasing their performance in individual events.
- **Nations without Gold Medals:** Countries that have never won gold medals but have won silver or bronze medals were identified, highlighting their achievements despite not achieving gold.

Conclusion

The analysis of Olympic Games data provides valuable insights into the history, participation, and achievements of one of the world's most celebrated sporting events. Through comprehensive SQL data analysis, we have uncovered a multitude of findings that shed light on the diverse and dynamic nature of the Olympic Games. Here are the key takeaways from this project:

- **Historical Context:** The Olympic Games have a rich history spanning over a century, with numerous editions held across various cities worldwide. Our analysis has provided a detailed overview of all Olympic Games held to date, including their years, seasons (Summer or Winter), and host cities.
- **Global Participation:** The participation of nations in the Olympic Games underscores the event's status as a truly global phenomenon. By analyzing the total number of participating nations in each edition, we have gained insights into the evolving landscape of international representation in the Games.
- **Variability in Participation:** The analysis has revealed significant variability in the number of participating countries across different editions of the Olympic Games. We identified years with both the highest and lowest participation rates, reflecting the dynamic nature of global sporting events.
- **Enduring Commitment:** Despite fluctuations in participation, certain nations have demonstrated an enduring commitment to the Olympic Games by participating in every edition. This highlights the event's universal appeal and the longstanding tradition of athletic excellence.

Conclusion

Sporting Diversity: The Olympic Games are a showcase of sporting diversity, featuring a wide array of athletic disciplines. Our analysis identified sports that have been played consistently in all Summer Olympic Games, as well as those that were featured only once, adding to the event's rich tapestry of competition.

Exceptional Achievements: The analysis unearthed remarkable achievements, such as the oldest athletes to win gold medals, showcasing the enduring spirit of athleticism and determination.

Gender Representation: We examined the gender ratio of athletes participating in Olympic Games, contributing to discussions on gender equity and representation in sports.

Athletic Excellence: The project highlighted the achievements of top-performing athletes, including those who have won the most gold medals in Olympic history, underscoring their unparalleled contributions to the sporting world.

National Success: We identified the top-performing countries in terms of total medals won, showcasing their dominance and success on the Olympic stage.

Medal Distribution: By examining the total number of gold, silver, and bronze medals won by each country, we gained insights into the distribution of medals and national achievements.

In conclusion, the analysis of Olympic Games data reaffirms the event's status as a symbol of unity, diversity, and sporting excellence on the global stage. Through meticulous data analysis, we have deepened our understanding of the Olympic Games' impact, legacy, and enduring significance in the world of sports.



THANK YOU

Thank you for your time, attention, and enthusiasm for the world of analytics!

