# Data Validation

## Olympics Athletes Web Scraping Project | Analysis Project (1896 - 2022)

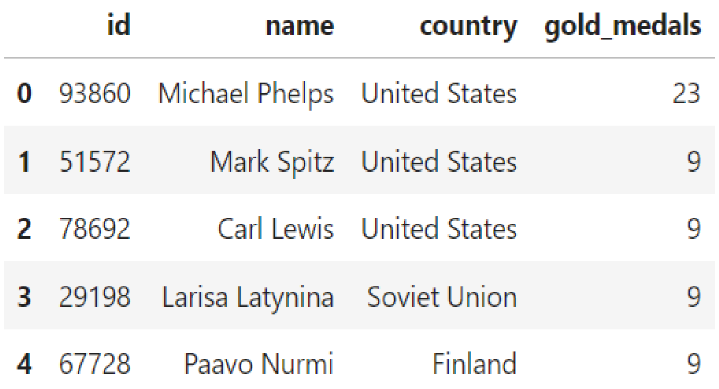
In this document, we validate our web-scraped data's accuracy by comparing it to Olympedia website and our PostgreSQL query outputs side-by-side. We have actually selected specific data points to ensure the accuracy of our data, ensuring its trustworthiness for analysis and engineering projects, and alignment with the Olympedia website.

Author: Ronnie Chan
Last modified: 2023/SEPT/5

### 13. Fetch the top 5 athletes who have won the most gold medals.

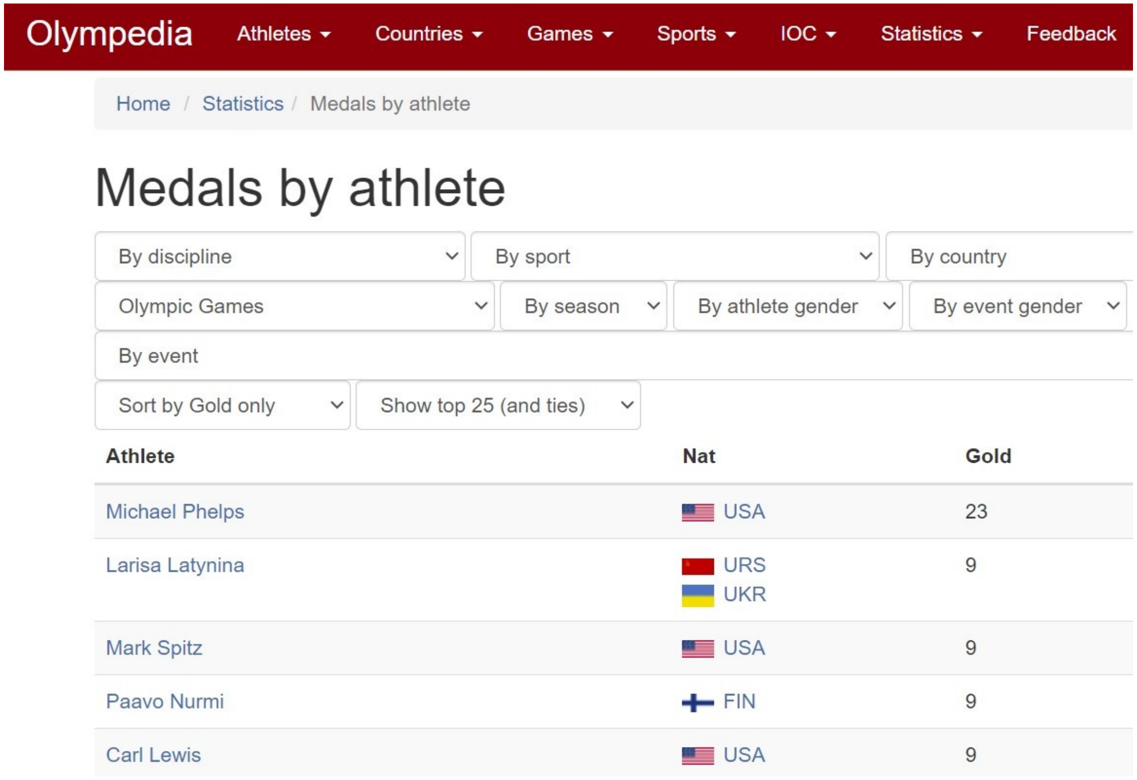
```
pd.read_sql("""
SELECT
    h.id,
    h.name,
    n.country,
    COUNT(h.medal) AS gold_medals
FROM olympics.olympic_history_cleaned h
JOIN olympics.noc_countries n ON h.noc = n.noc
WHERE h.event LIKE '%Olympic%' AND h.medal = 'Gold'
GROUP BY h.id, h.name, n.country, h.noc
ORDER BY gold_medals DESC, n.country DESC
LIMIT 5
""", conn)
```

Here is my SQL output:



|   | id | name | country | gold_medals |
|---|---|---|---|---|
| **0** | 93860 | Michael Phelps | United States | 23 |
| **1** | 51572 | Mark Spitz | United States | 9 |
| **2** | 78692 | Carl Lewis | United States | 9 |
| **3** | 29198 | Larisa Latynina | Soviet Union | 9 |
| **4** | 67728 | Paavo Nurmi | Finland | 9 |

Here is the Olympedia data:



Home / Statistics / Medals by athlete

# Medals by athlete

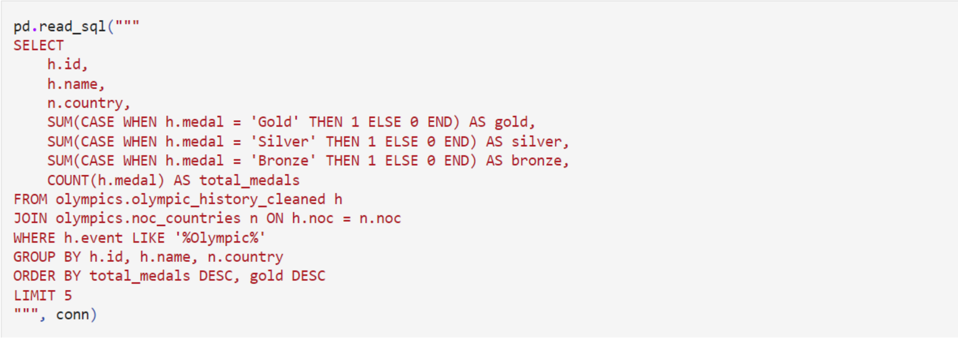
Sort by Gold only

Show top 25 (and ties)

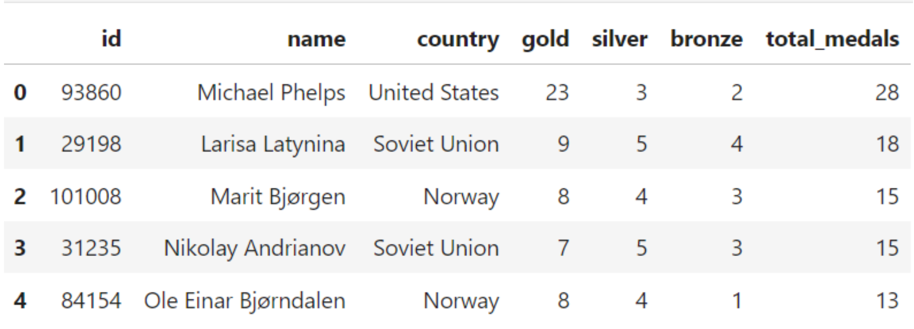
| Athlete | Nat | Gold |
|---|---|---|
| Michael Phelps | USA | 23 |
| Larisa Latynina | URS<br>UKR | 9 |
| Mark Spitz | USA | 9 |
| Paavo Nurmi | FIN | 9 |
| Carl Lewis | USA | 9 |

### 14. Fetch the top 5 athletes who have won the most medals (gold/silver/bronze).

Here is my SQL query and its corresponding output.

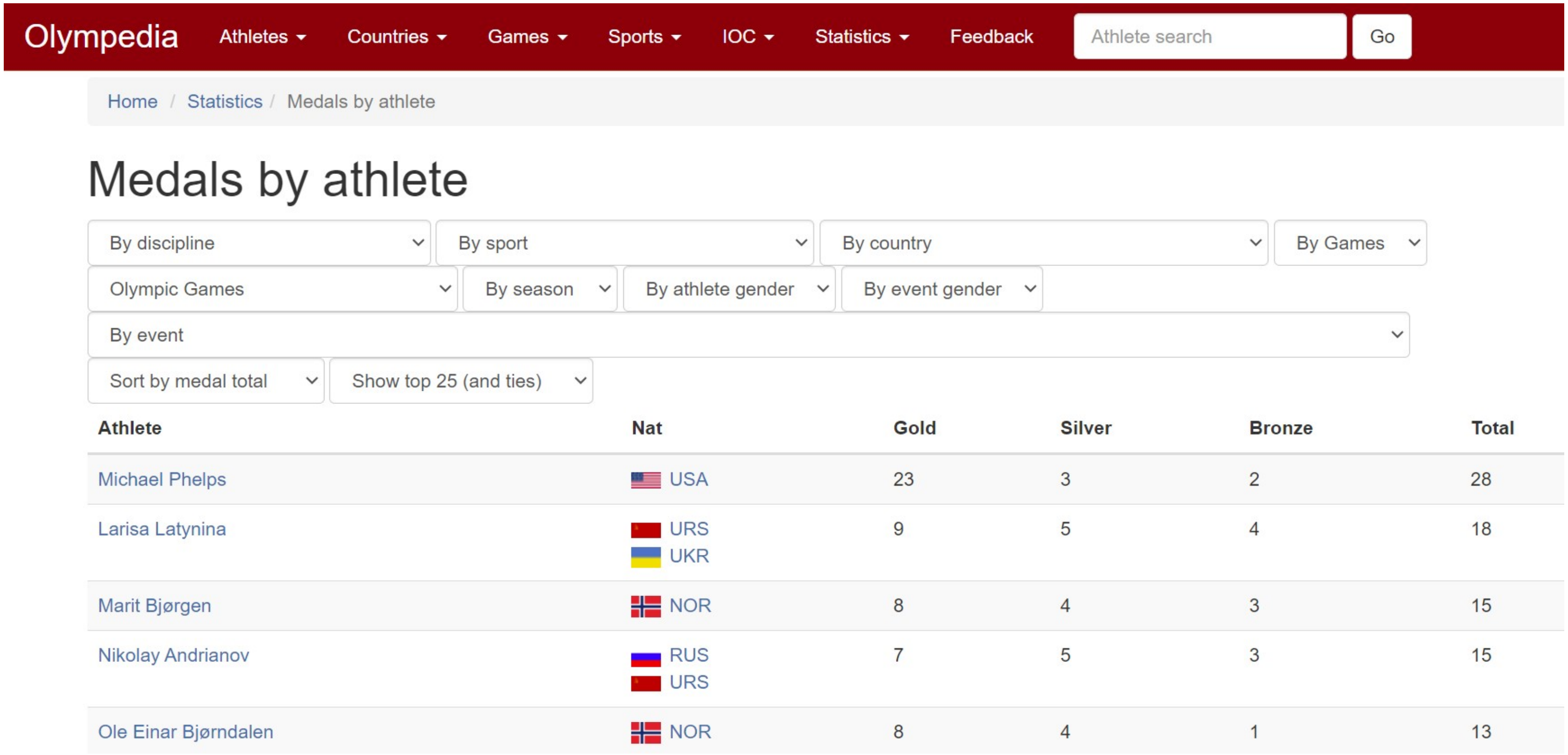


```
pd.read_sql("""
SELECT
    h.id,
    h.name,
    n.country,
    SUM(CASE WHEN h.medal = 'Gold' THEN 1 ELSE 0 END) AS gold,
    SUM(CASE WHEN h.medal = 'Silver' THEN 1 ELSE 0 END) AS silver,
    SUM(CASE WHEN h.medal = 'Bronze' THEN 1 ELSE 0 END) AS bronze,
    COUNT(h.medal) AS total_medals
FROM olympics.olympic_history_cleaned h
JOIN olympics.noc_countries n ON h.noc = n.noc
WHERE h.event LIKE '%Olympic%'
GROUP BY h.id, h.name, n.country
ORDER BY total_medals DESC, gold DESC
LIMIT 5
""", conn)
```



| | id | name | country | gold | silver | bronze | total_medals |
|---|---|---|---|---|---|---|---|
| **0** | 93860 | Michael Phelps | United States | 23 | 3 | 2 | 28 |
| **1** | 29198 | Larisa Latynina | Soviet Union | 9 | 5 | 4 | 18 |
| **2** | 101008 | Marit Bjørgen | Norway | 8 | 4 | 3 | 15 |
| **3** | 31235 | Nikolay Andrianov | Soviet Union | 7 | 5 | 3 | 15 |
| **4** | 84154 | Ole Einar Bjørndalen | Norway | 8 | 4 | 1 | 13 |

Here is the Olympedia data:



Home / Statistics / Medals by athlete

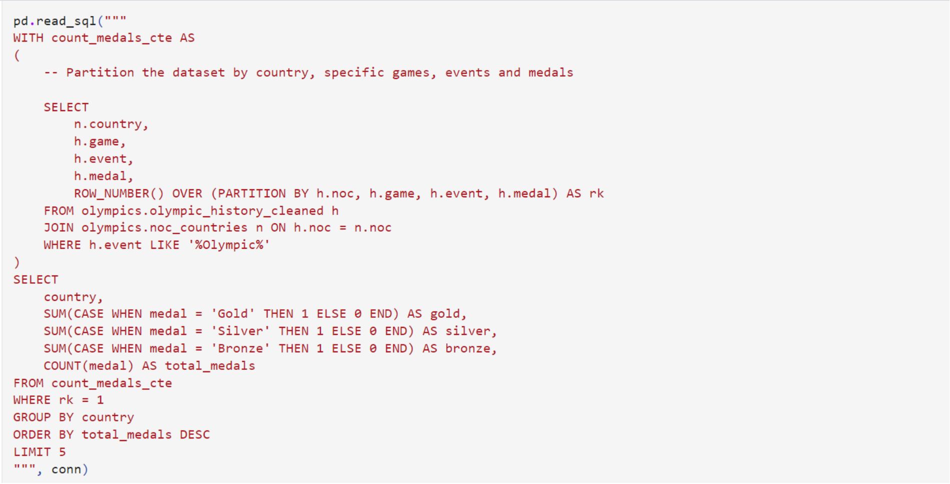
Medals by athlete

By discipline | By sport | By country | By Games
Olympic Games | By season | By athlete gender | By event gender
By event
Sort by medal total | Show top 25 (and ties)

| Athlete | Nat | Gold | Silver | Bronze | Total |
|---|---|---|---|---|---|
| Michael Phelps | USA | 23 | 3 | 2 | 28 |
| Larisa Latynina | URS<br>UKR | 9 | 5 | 4 | 18 |
| Marit Bjørgen | NOR | 8 | 4 | 3 | 15 |
| Nikolay Andrianov | RUS<br>URS | 7 | 5 | 3 | 15 |
| Ole Einar Bjørndalen | NOR | 8 | 4 | 1 | 13 |

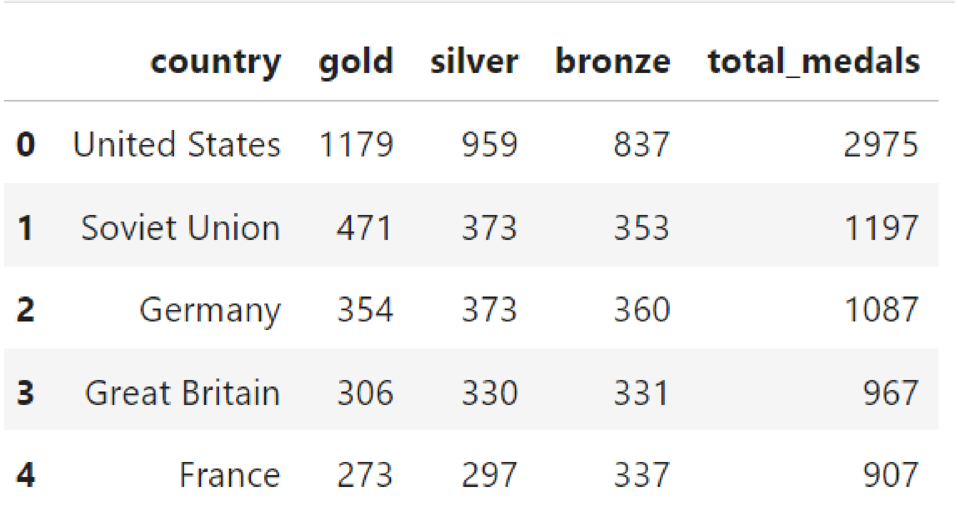
### 16. List down total gold, silver and bronze medals won by each country.

Here is my SQL query and its corresponding output.

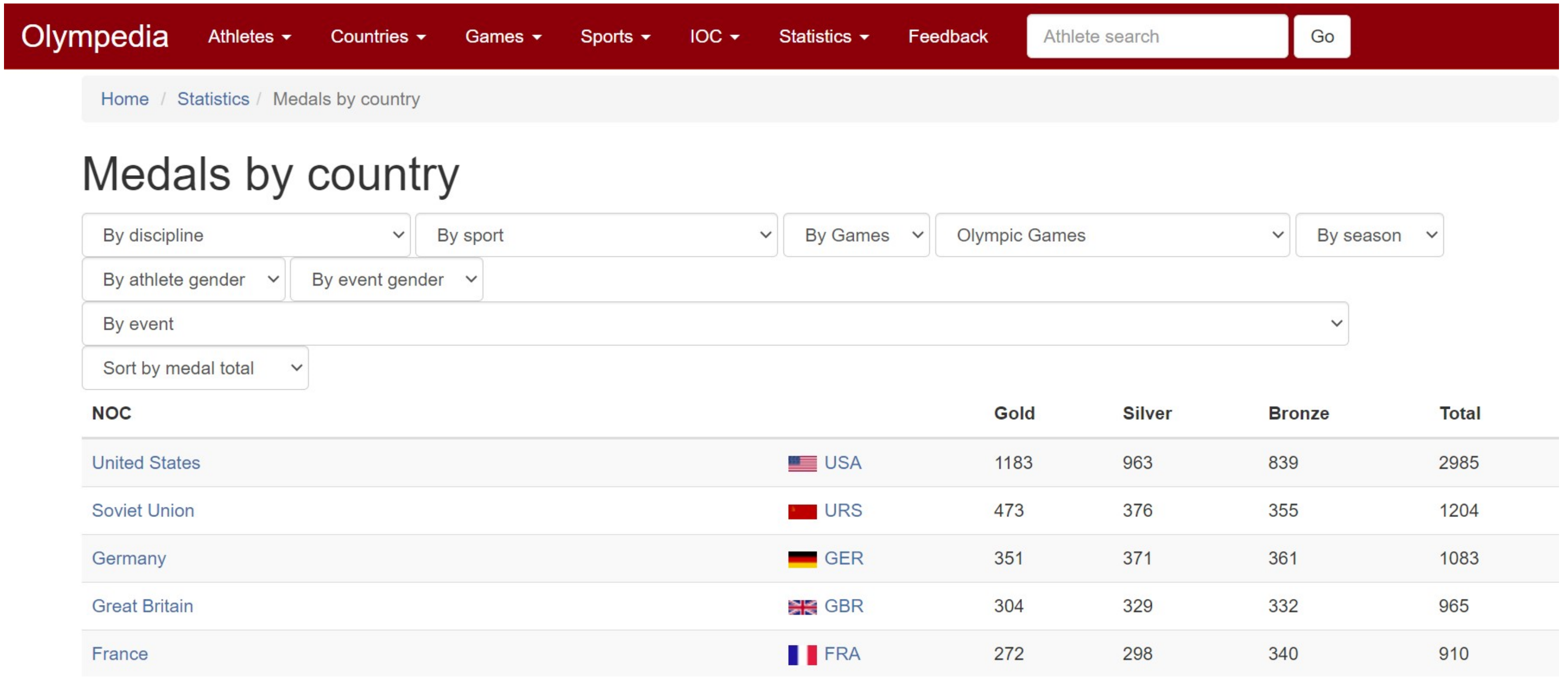


```
pd.read_sql("""
WITH count_medals_cte AS
(
    -- Partition the dataset by country, specific games, events and medals

    SELECT
        n.country,
        h.game,
        h.event,
        h.medal,
        ROW_NUMBER() OVER (PARTITION BY h.noc, h.game, h.event, h.medal) AS rk
    FROM olympics.olympic_history_cleaned h
    JOIN olympics.noc_countries n ON h.noc = n.noc
    WHERE h.event LIKE '%Olympic%'
)
SELECT
    country,
    SUM(CASE WHEN medal = 'Gold' THEN 1 ELSE 0 END) AS gold,
    SUM(CASE WHEN medal = 'Silver' THEN 1 ELSE 0 END) AS silver,
    SUM(CASE WHEN medal = 'Bronze' THEN 1 ELSE 0 END) AS bronze,
    COUNT(medal) AS total_medals
FROM count_medals_cte
WHERE rk = 1
GROUP BY country
ORDER BY total_medals DESC
LIMIT 5
""", conn)
```



| | country | gold | silver | bronze | total_medals |
|---|---|---|---|---|---|
| 0 | United States | 1179 | 959 | 837 | 2975 |
| 1 | Soviet Union | 471 | 373 | 353 | 1197 |
| 2 | Germany | 354 | 373 | 360 | 1087 |
| 3 | Great Britain | 306 | 330 | 331 | 967 |
| 4 | France | 273 | 297 | 337 | 907 |



Medals by country

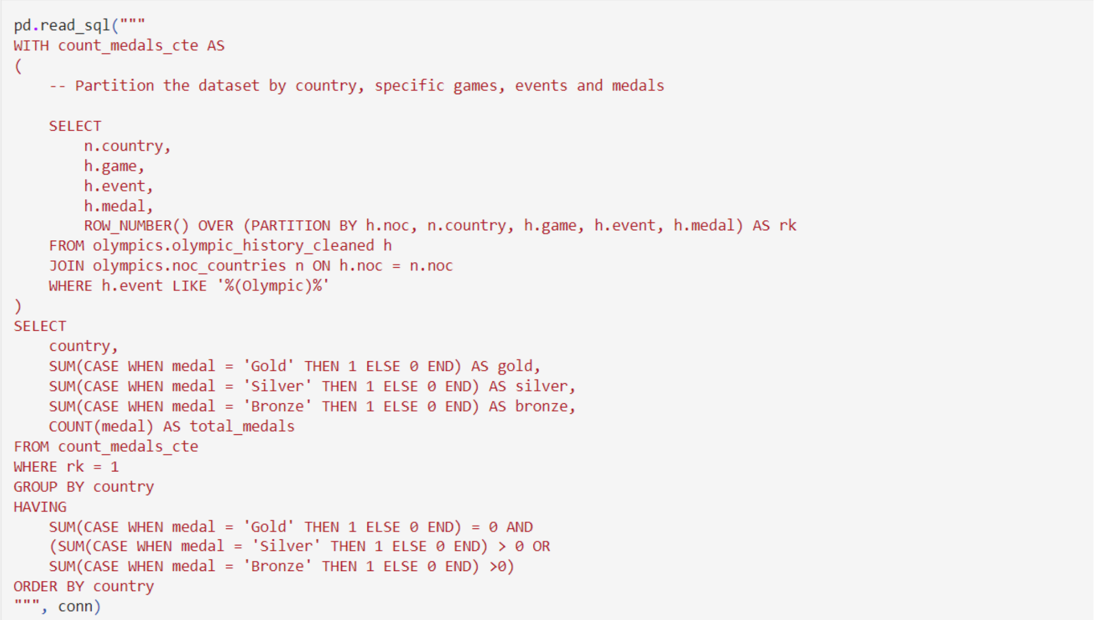
By discipline | By sport | By Games | Olympic Games | By season
By athlete gender | By event gender
By event
Sort by medal total

| NOC | | Gold | Silver | Bronze | Total |
|---|---|---|---|---|---|
| United States | USA | 1183 | 963 | 839 | 2985 |
| Soviet Union | URS | 473 | 376 | 355 | 1204 |
| Germany | GER | 351 | 371 | 361 | 1083 |
| Great Britain | GBR | 304 | 329 | 332 | 965 |
| France | FRA | 272 | 298 | 340 | 910 |

**Note:** The data shows a slight deviation of approximately 10 units from the expected values; however, this variance still confirms the overall validity of the dataset.

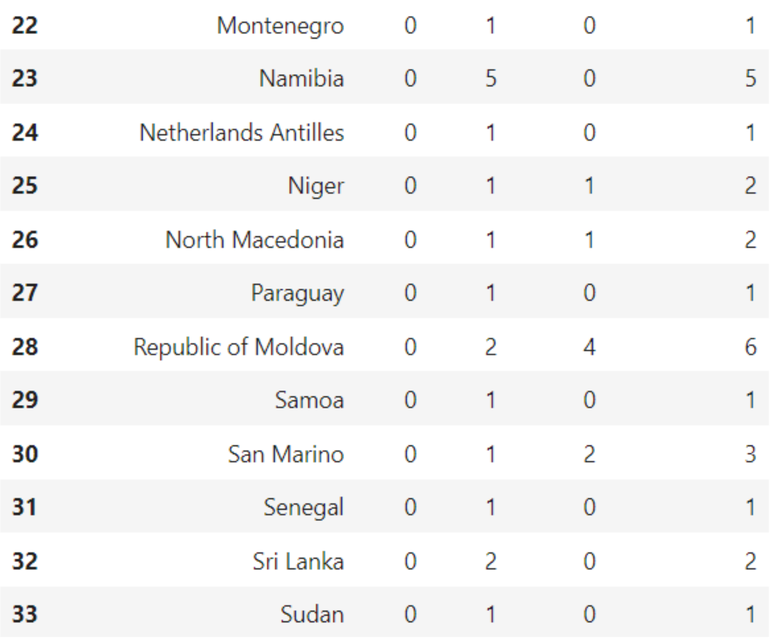
### 19. Which countries have never won a gold medal but have won silver/bronze medals?

Here is my SQL query and a portion of its corresponding output.

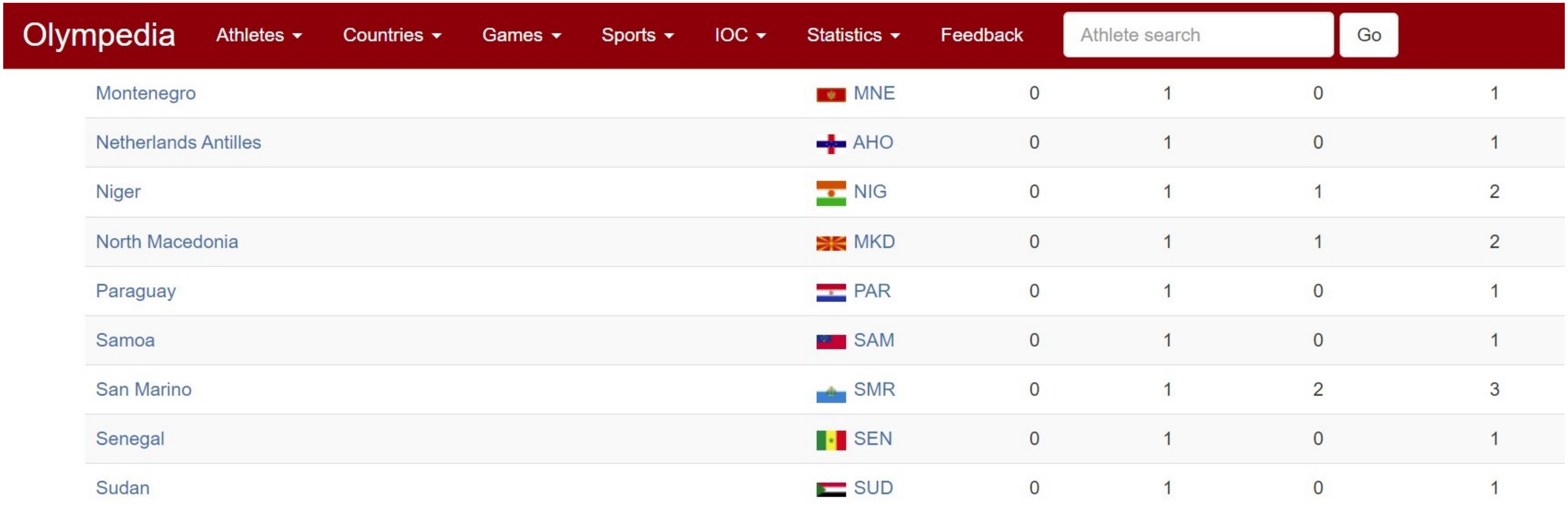


```
pd.read_sql("""
WITH count_medals_cte AS
(
    -- Partition the dataset by country, specific games, events and medals

    SELECT
        n.country,
        h.game,
        h.event,
        h.medal,
        ROW_NUMBER() OVER (PARTITION BY h.noc, n.country, h.game, h.event, h.medal) AS rk
    FROM olympics.olympic_history_cleaned h
    JOIN olympics.noc_countries n ON h.noc = n.noc
    WHERE h.event LIKE '%(Olympic)%'
)
SELECT
    country,
    SUM(CASE WHEN medal = 'Gold' THEN 1 ELSE 0 END) AS gold,
    SUM(CASE WHEN medal = 'Silver' THEN 1 ELSE 0 END) AS silver,
    SUM(CASE WHEN medal = 'Bronze' THEN 1 ELSE 0 END) AS bronze,
    COUNT(medal) AS total_medals
FROM count_medals_cte
WHERE rk = 1
GROUP BY country
HAVING
    SUM(CASE WHEN medal = 'Gold' THEN 1 ELSE 0 END) = 0 AND
    (SUM(CASE WHEN medal = 'Silver' THEN 1 ELSE 0 END) > 0 OR
    SUM(CASE WHEN medal = 'Bronze' THEN 1 ELSE 0 END) >0)
ORDER BY country
""", conn)
```



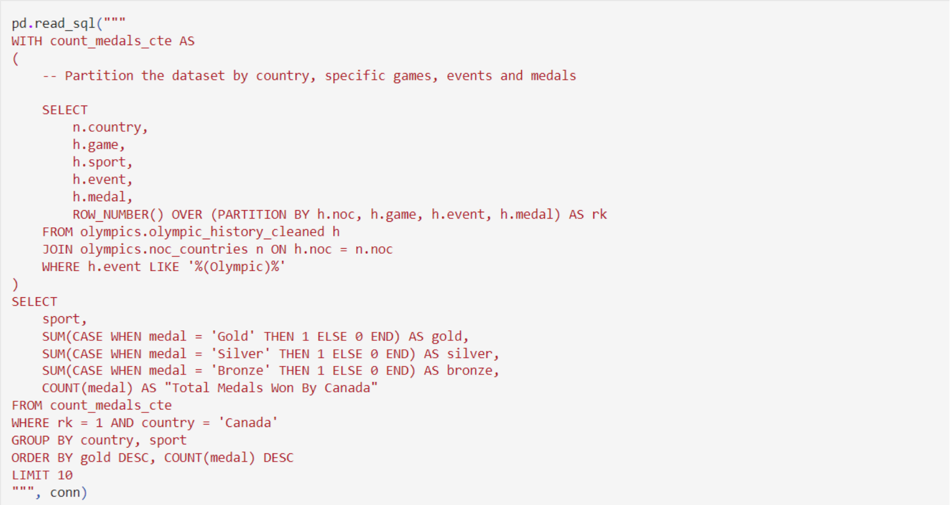
| | | | | | |
|---|---|---|---|---|---|
| **22** | Montenegro | 0 | 1 | 0 | 1 |
| **23** | Namibia | 0 | 5 | 0 | 5 |
| **24** | Netherlands Antilles | 0 | 1 | 0 | 1 |
| **25** | Niger | 0 | 1 | 1 | 2 |
| **26** | North Macedonia | 0 | 1 | 1 | 2 |
| **27** | Paraguay | 0 | 1 | 0 | 1 |
| **28** | Republic of Moldova | 0 | 2 | 4 | 6 |
| **29** | Samoa | 0 | 1 | 0 | 1 |
| **30** | San Marino | 0 | 1 | 2 | 3 |
| **31** | Senegal | 0 | 1 | 0 | 1 |
| **32** | Sri Lanka | 0 | 2 | 0 | 2 |
| **33** | Sudan | 0 | 1 | 0 | 1 |



| | | | | | |
|---|---|---|---|---|---|
| Montenegro | MNE | 0 | 1 | 0 | 1 |
| Netherlands Antilles | AHO | 0 | 1 | 0 | 1 |
| Niger | NIG | 0 | 1 | 1 | 2 |
| North Macedonia | MKD | 0 | 1 | 1 | 2 |
| Paraguay | PAR | 0 | 1 | 0 | 1 |
| Samoa | SAM | 0 | 1 | 0 | 1 |
| San Marino | SMR | 0 | 1 | 2 | 3 |
| Senegal | SEN | 0 | 1 | 0 | 1 |
| Sudan | SUD | 0 | 1 | 0 | 1 |

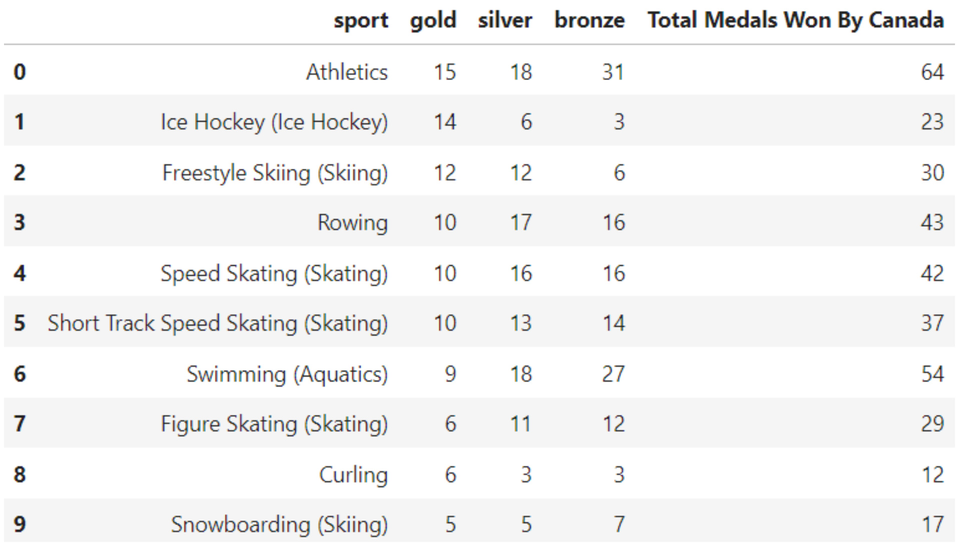
### 20. In which sport/event Canada has won the highest number of medals.

Here is my SQL query and its corresponding output.

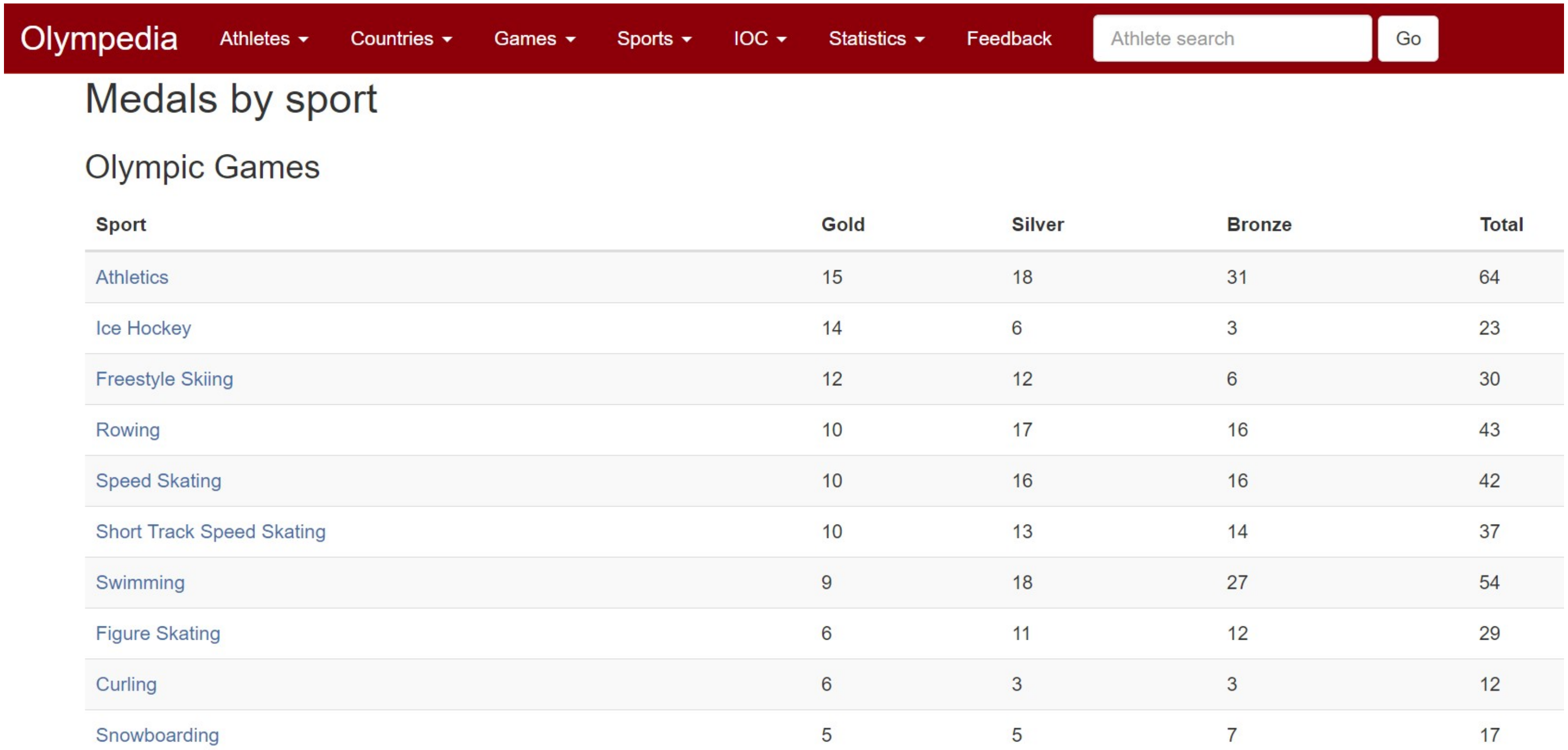


```
pd.read_sql("""
WITH count_medals_cte AS
(
    -- Partition the dataset by country, specific games, events and medals

    SELECT
        n.country,
        h.game,
        h.sport,
        h.event,
        h.medal,
        ROW_NUMBER() OVER (PARTITION BY h.noc, h.game, h.event, h.medal) AS rk
    FROM olympics.olympic_history_cleaned h
    JOIN olympics.noc_countries n ON h.noc = n.noc
    WHERE h.event LIKE '%(Olympic)%'
)
SELECT
    sport,
    SUM(CASE WHEN medal = 'Gold' THEN 1 ELSE 0 END) AS gold,
    SUM(CASE WHEN medal = 'Silver' THEN 1 ELSE 0 END) AS silver,
    SUM(CASE WHEN medal = 'Bronze' THEN 1 ELSE 0 END) AS bronze,
    COUNT(medal) AS "Total Medals Won By Canada"
FROM count_medals_cte
WHERE rk = 1 AND country = 'Canada'
GROUP BY country, sport
ORDER BY gold DESC, COUNT(medal) DESC
LIMIT 10
""", conn)
```



|  | sport | gold | silver | bronze | Total Medals Won By Canada |
|---|---|---|---|---|---|
| **0** | Athletics | 15 | 18 | 31 | 64 |
| **1** | Ice Hockey (Ice Hockey) | 14 | 6 | 3 | 23 |
| **2** | Freestyle Skiing (Skiing) | 12 | 12 | 6 | 30 |
| **3** | Rowing | 10 | 17 | 16 | 43 |
| **4** | Speed Skating (Skating) | 10 | 16 | 16 | 42 |
| **5** | Short Track Speed Skating (Skating) | 10 | 13 | 14 | 37 |
| **6** | Swimming (Aquatics) | 9 | 18 | 27 | 54 |
| **7** | Figure Skating (Skating) | 6 | 11 | 12 | 29 |
| **8** | Curling | 6 | 3 | 3 | 12 |
| **9** | Snowboarding (Skiing) | 5 | 5 | 7 | 17 |



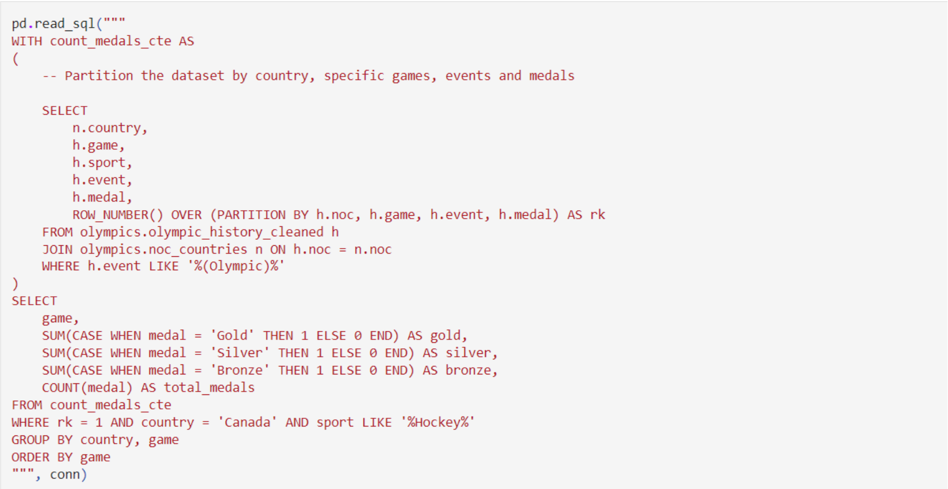
## Medals by sport

### Olympic Games

| Sport | Gold | Silver | Bronze | Total |
|---|---|---|---|---|
| Athletics | 15 | 18 | 31 | 64 |
| Ice Hockey | 14 | 6 | 3 | 23 |
| Freestyle Skiing | 12 | 12 | 6 | 30 |
| Rowing | 10 | 17 | 16 | 43 |
| Speed Skating | 10 | 16 | 16 | 42 |
| Short Track Speed Skating | 10 | 13 | 14 | 37 |
| Swimming | 9 | 18 | 27 | 54 |
| Figure Skating | 6 | 11 | 12 | 29 |
| Curling | 6 | 3 | 3 | 12 |
| Snowboarding | 5 | 5 | 7 | 17 |

## 21. Break down all olympic games where Canada won medals for Hockey and how many medals in each olympic games.

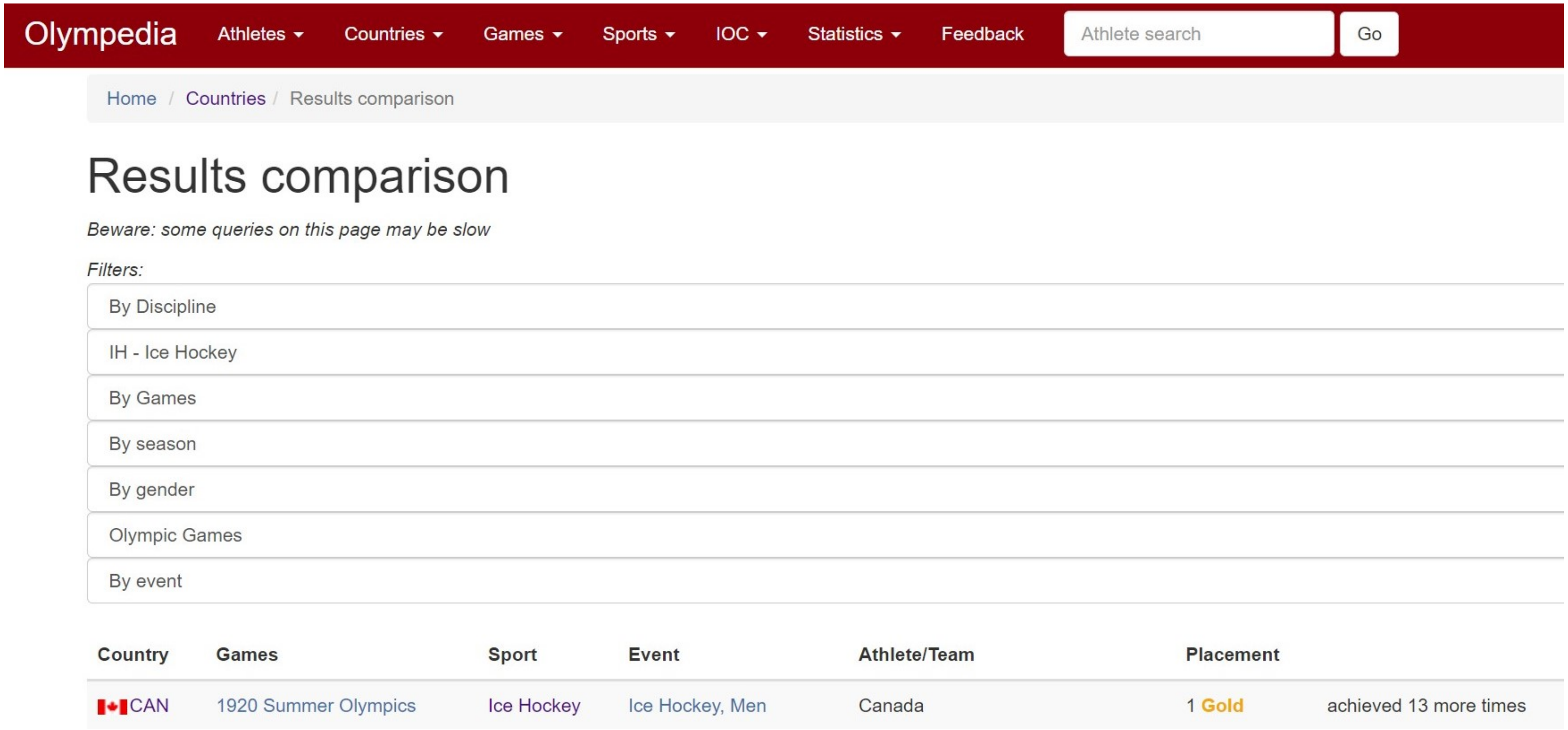
Here is my SQL query and the piece of its corresponding output.



```
pd.read_sql("""
WITH count_medals_cte AS
(
    -- Partition the dataset by country, specific games, events and medals

    SELECT
        n.country,
        h.game,
        h.sport,
        h.event,
        h.medal,
        ROW_NUMBER() OVER (PARTITION BY h.noc, h.game, h.event, h.medal) AS rk
    FROM olympics.olympic_history_cleaned h
    JOIN olympics.noc_countries n ON h.noc = n.noc
    WHERE h.event LIKE '%(Olympic)%'
)
SELECT
    game,
    SUM(CASE WHEN medal = 'Gold' THEN 1 ELSE 0 END) AS gold,
    SUM(CASE WHEN medal = 'Silver' THEN 1 ELSE 0 END) AS silver,
    SUM(CASE WHEN medal = 'Bronze' THEN 1 ELSE 0 END) AS bronze,
    COUNT(medal) AS total_medals
FROM count_medals_cte
WHERE rk = 1 AND country = 'Canada' AND sport LIKE '%Hockey%'
GROUP BY country, game
ORDER BY game
""", conn)
```

In my query output, Canada has indeed won 14 gold medals in all-time Olympic games. To view the complete dataset, please refer to the 'olympics_analysis' notebook. This validation confirms the accuracy of my scraping technique.



# Results comparison

*Beware: some queries on this page may be slow*

*Filters:*

By Discipline

IH - Ice Hockey

By Games

By season

By gender

Olympic Games

By event

| Country | Games | Sport | Event | Athlete/Team | Placement | |
|---|---|---|---|---|---|---|
| CAN | 1920 Summer Olympics | Ice Hockey | Ice Hockey, Men | Canada | 1 Gold | achieved 13 more times |

# Conclusion

After comparing my query results to the data on Olympedia, it’s clear that my data scraping process was accurate. The two datasets match up well with only minor differences, showing that the information I collected is reliable. This validation gives us confidence in the quality of our data, which is crucial for our future analysis and projects.