Preparing for Project Proposal

Which Client and Why?

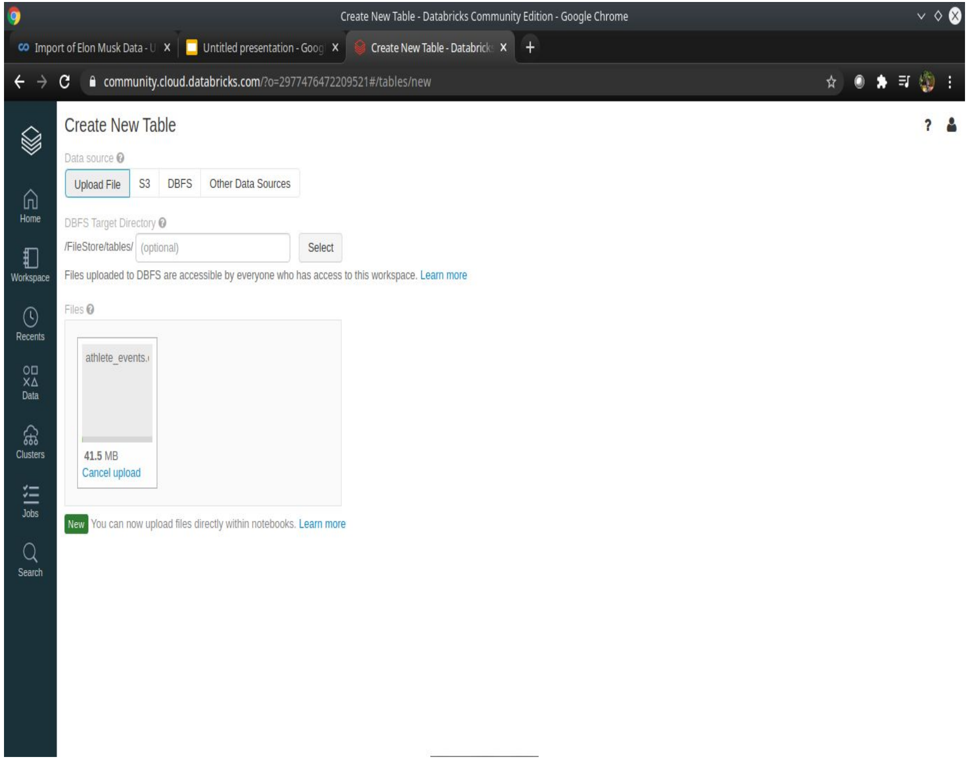
For this Milestone, I selected the Client 3, that is, **SportsStats.**

This client, after the analysis, can help many trainers and the authorities in training their athletes so that they can perform better in their respective sports. This in turn will result in much tougher competition between the athletes and hence more entertainment for the audience. Thus, a WIN-WIN situation for all.

Importing the Data

I chose to work with Databricks. Hence, I took the following steps to import the data:

- Downloaded the zipped csv files from **Dropbox** and un-zipped them.
- Signing in to **Databricks Community edition**, created a **new cluster**.
- From the **DATA** option, opter for ‘**Add Data**’ and uploaded the csv files.
- After previewing the table, selected ‘**Create Table**’ and DONE.



Create New Table - Databricks Community Edition - Google Chrome

community.cloud.databricks.com/?o=2977476472209521#/tables/new

Create New Table

Data source

DBFS Target Directory

/FileStore/tables/ (optional)

Files uploaded to DBFS are accessible by everyone who has access to this workspace. Learn more

Files

athlete_events.

41.5 MB

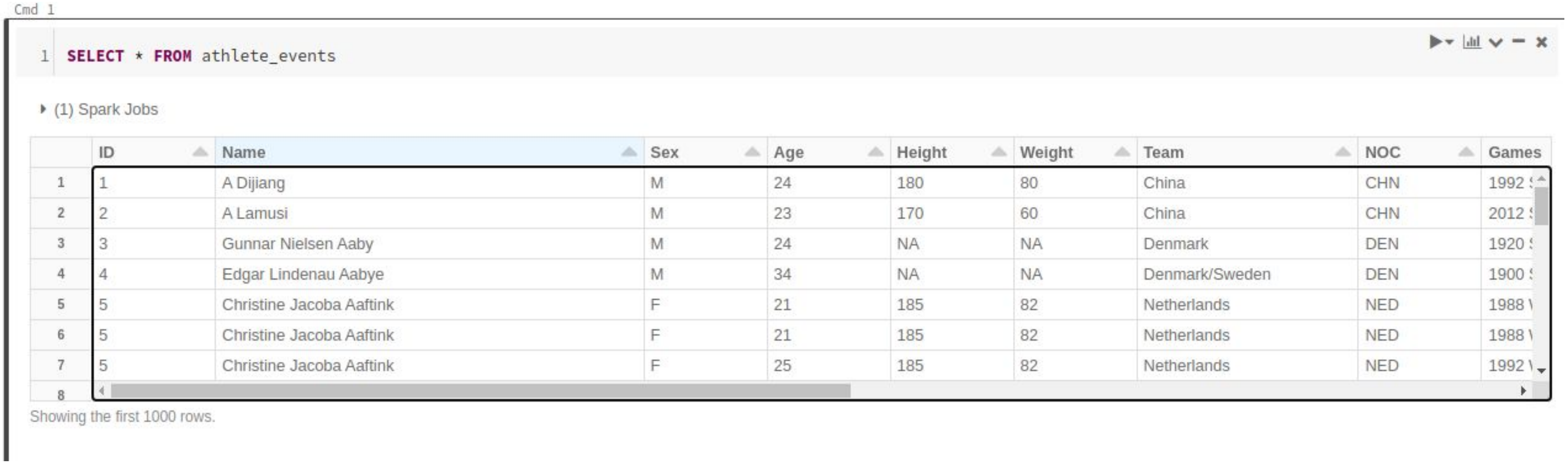
Cancel upload

New You can now upload files directly within notebooks. Learn more

Import of Elon Musk Data - U
Untitled presentation - Goog
Create New Table - Databricks
Home
Workspace
Recents
Data
Clusters
Jobs
Search
Upload File
S3
DBFS
Other Data Sources
Select

Some Stats

- Viewing all the records:



Cmd 1

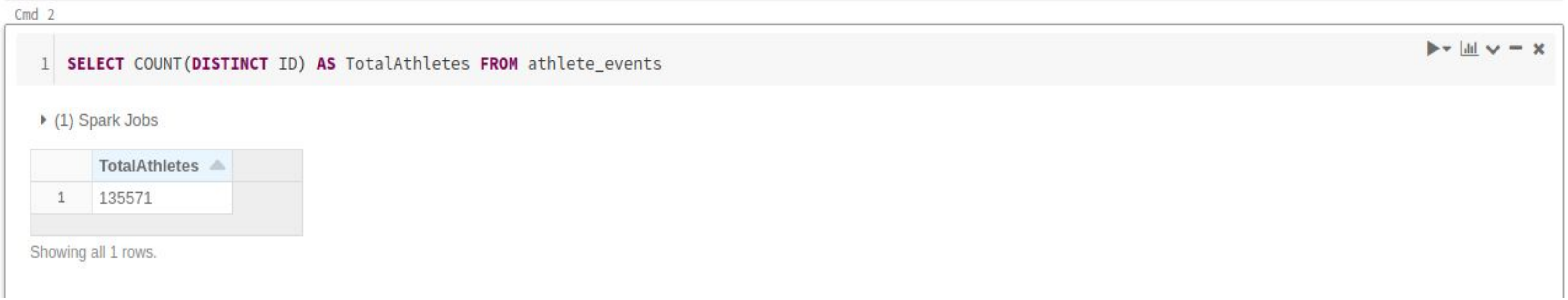
```
SELECT * FROM athlete_events
```

(1) Spark Jobs

	ID	Name	Sex	Age	Height	Weight	Team	NOC	Games
1	1	A Dijiang	M	24	180	80	China	CHN	1992 S
2	2	A Lamusi	M	23	170	60	China	CHN	2012 S
3	3	Gunnar Nielsen Aaby	M	24	NA	NA	Denmark	DEN	1920 S
4	4	Edgar Lindenau Aabye	M	34	NA	NA	Denmark/Sweden	DEN	1900 S
5	5	Christine Jacoba Aaftink	F	21	185	82	Netherlands	NED	1988 W
6	5	Christine Jacoba Aaftink	F	21	185	82	Netherlands	NED	1988 W
7	5	Christine Jacoba Aaftink	F	25	185	82	Netherlands	NED	1992 W
8									

Showing the first 1000 rows.

- Counting the number of athletes in the dataset, we find that we have info on 135571 athletes.



Cmd 2

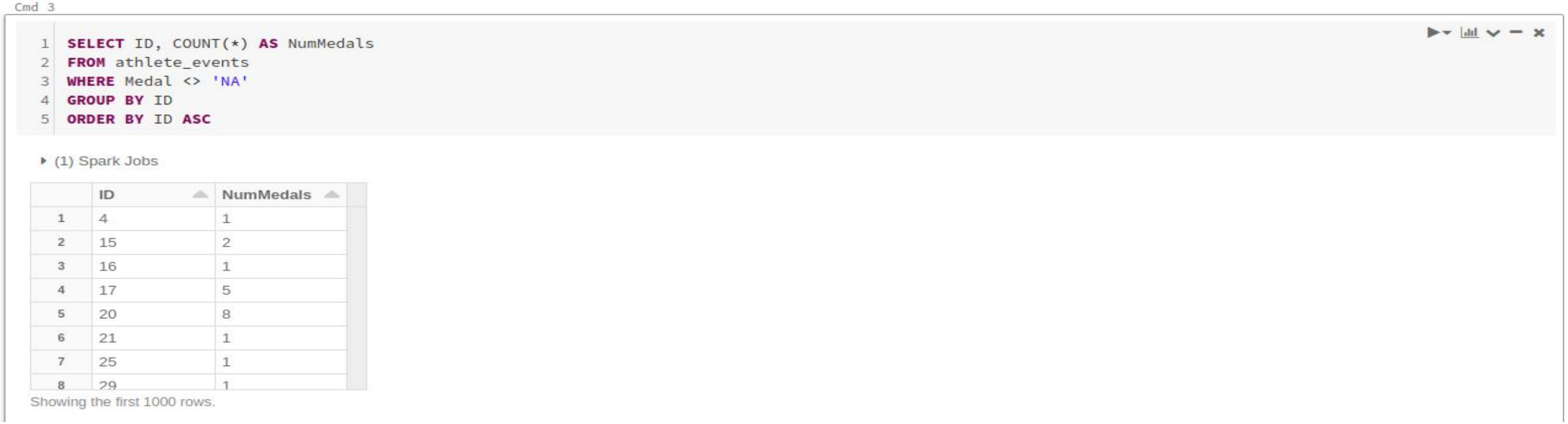
```sql
SELECT COUNT(DISTINCT ID) AS TotalAthletes FROM athlete_events
```

(1) Spark Jobs

	TotalAthletes
1	135571

Showing all 1 rows.

- Also we can find, how many of these athletes have won a medal.



Cmd 3

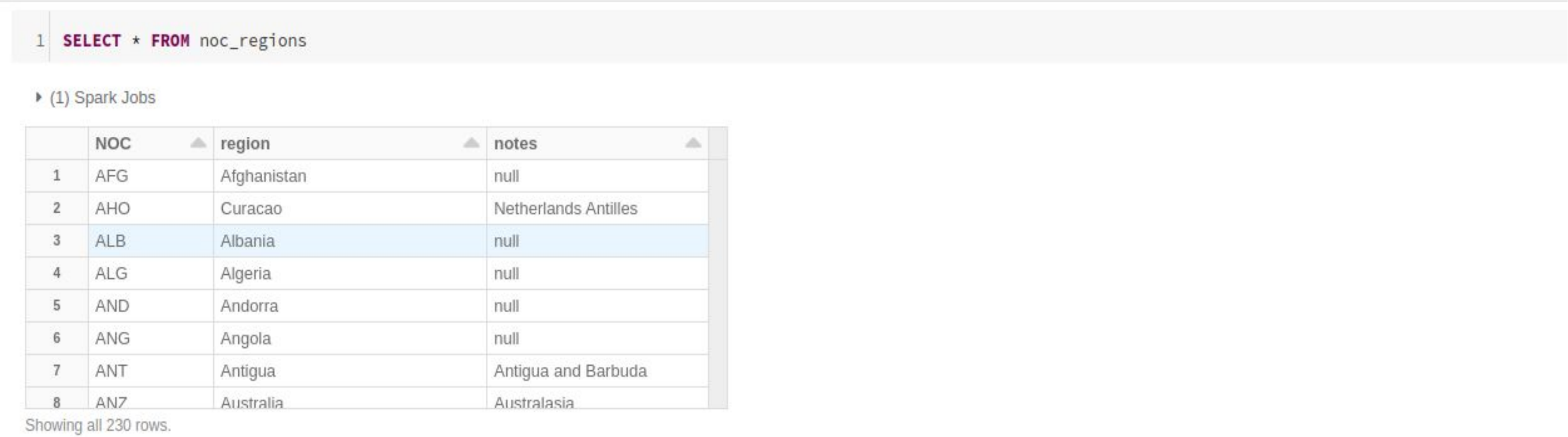
```sql
SELECT ID, COUNT(*) AS NumMedals
FROM athlete_events
WHERE Medal <> 'NA'
GROUP BY ID
ORDER BY ID ASC
```

(1) Spark Jobs

	ID	NumMedals
1	4	1
2	15	2
3	16	1
4	17	5
5	20	8
6	21	1
7	25	1
8	29	1

Showing the first 1000 rows.

- There are 230 regions as given by the noc_regions dataset.



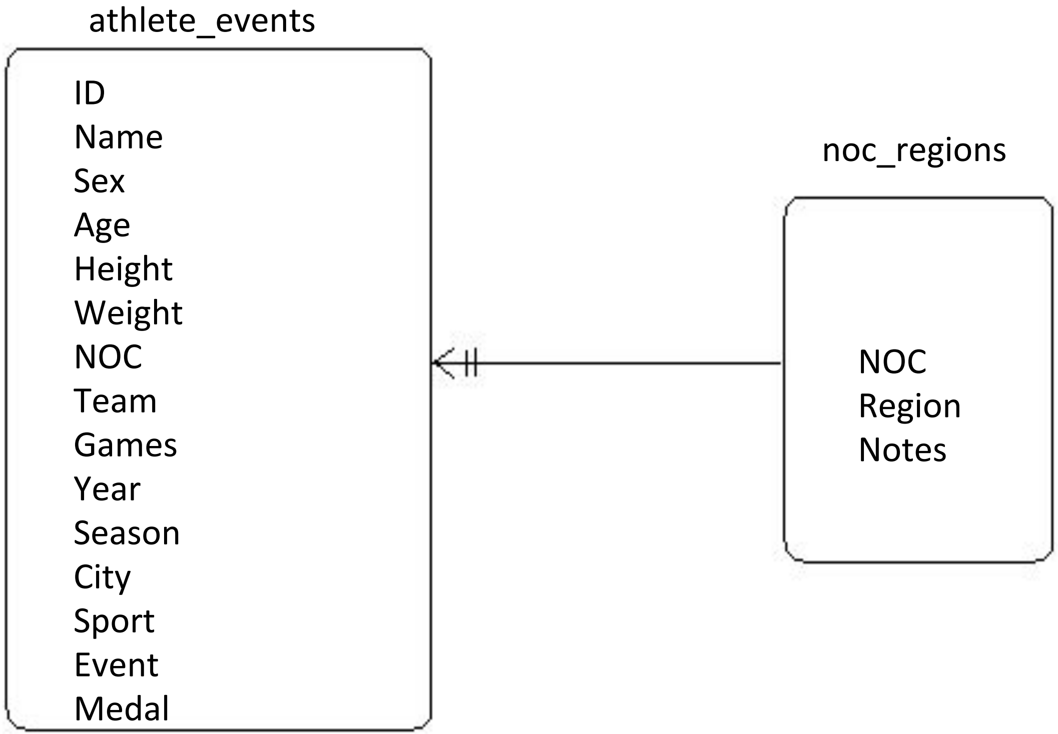
```sql
SELECT * FROM noc_regions
```

▸ (1) Spark Jobs

	NOC	region	notes
1	AFG	Afghanistan	null
2	AHO	Curacao	Netherlands Antilles
3	ALB	Albania	null
4	ALG	Algeria	null
5	AND	Andorra	null
6	ANG	Angola	null
7	ANT	Antigua	Antigua and Barbuda
8	ANZ	Australia	Australasia

Showing all 230 rows.

Entity Relationship Diagram (ERD)



athlete_events

ID
Name
Sex
Age
Height
Weight
NOC
Team
Games
Year
Season
City
Sport
Event
Medal

noc_regions

NOC
Region
Notes

Project Proposal

Data Analysis Project Proposal for SportsStats
October 2020
Akshu Chahar

Project Description / Audience

In this project, we use the datasets **athlete_events** and **noc_regions** to find out some relationships between the performances of athletes in different games and events at different age across the world.

AUDIENCE: This project majorly targets the sports authorities and the trainers who have supervision over the listed athletes. Through this project they might find out the patterns between the performances of an athlete over a period of time and have them train accordingly.

Major Questions

- Does age affect any athlete’s performance?
- If it does, is there any relationship?
- Does this relationship apply to everyone?

Hypothesis

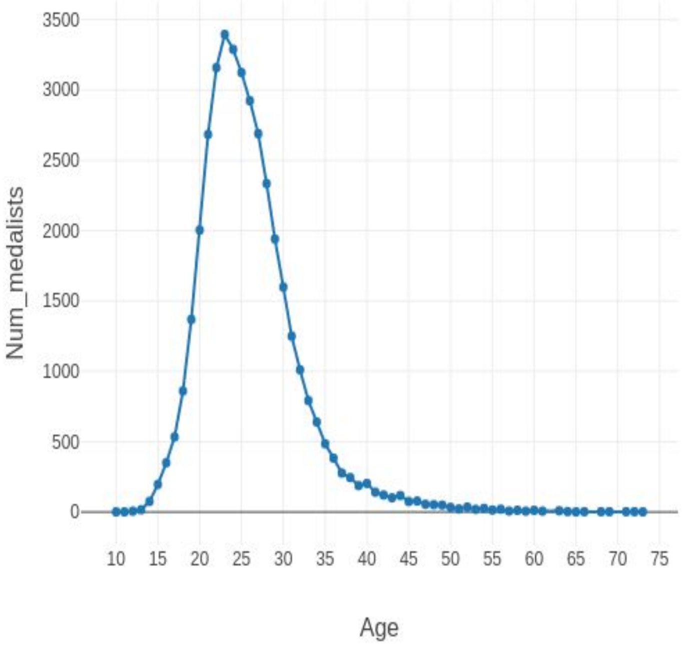
- Any athlete performs his/her during **24-28 years of age.**
- There should be a relationship between the age and performance of an athlete.
- This relationship should more or less apply to everyone.

Approach

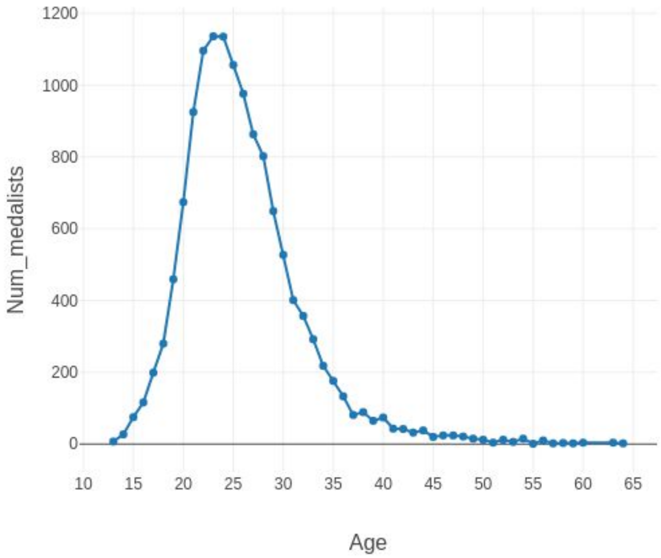
- Looking at the **Season** and **Medal** columns of **athlete_events**, we can find out if there is any relationship between the seasons and athlete's performance.
- Looking at the **Region** and **Medal** columns of **athlete_events**, we can find out if there is any relationship between the places and athlete's performance.
- Looking at the **Event** and **Medal** columns of **athlete_events**, we can find out if any particular event effects the athlete's performance.
- Looking at the **Age** and **Medal** columns of **athlete_events**, we can find out the relationship between the athlete's age and performance.

Descriptive Stats

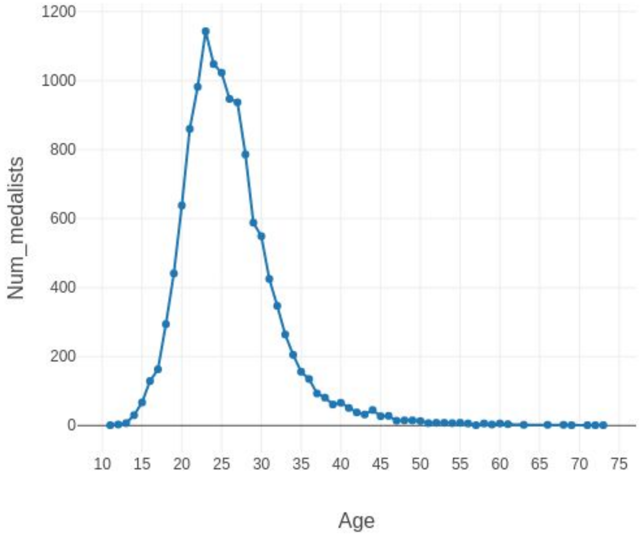
- We assumed that athletes are at their best during the age of 24 to 28. After performing some analysis, we can find out that most athletes won at the age of **23**.
- Moreover, athletes of age **20** to **29** have won more medals.



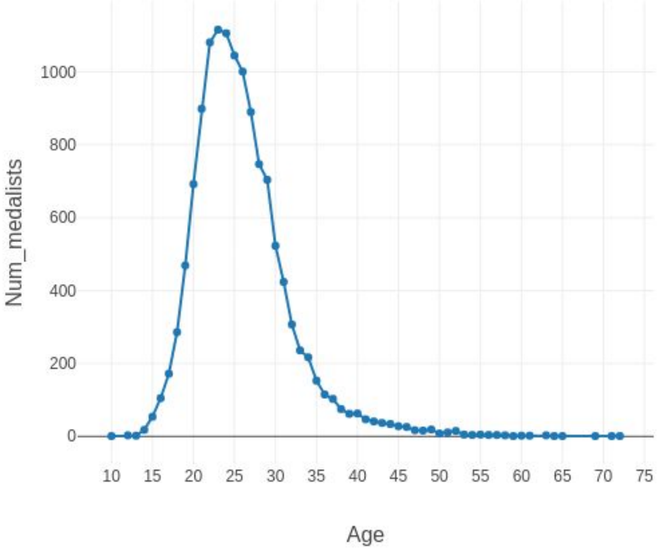
Apart from winning medals, it would be more interesting to find out the distribution of Gold, Silver and Bronze medals with respect to the athletes’ age.



- **<u>GOLD MEDAL</u>:** At the age of **23, 1136** athletes won Gold Medal followed by athletes of age **24** (**1135**). This in turns follows the overall trend.
- **<u>SILVER MEDAL</u>:** The highest number of Silver Medals are won by athletes of age **23 (1143)**. The range of winners, again follows the overall trend.
- **<u>BRONZE MEDAL</u>:** Athletes at the age of **23,** won **1116** bronze medals which is highest among all age groups.



SILVER MEDALISTS

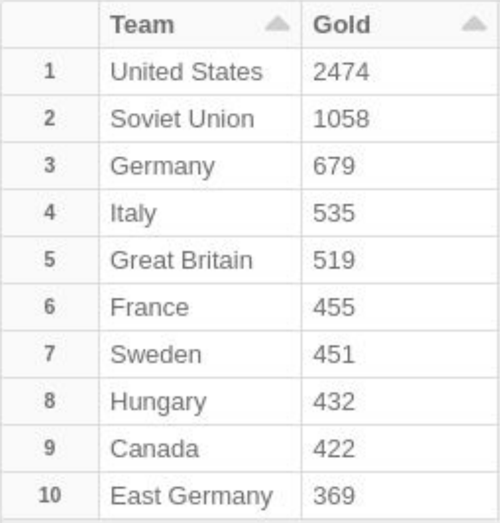


BRONZE MEDALISTS

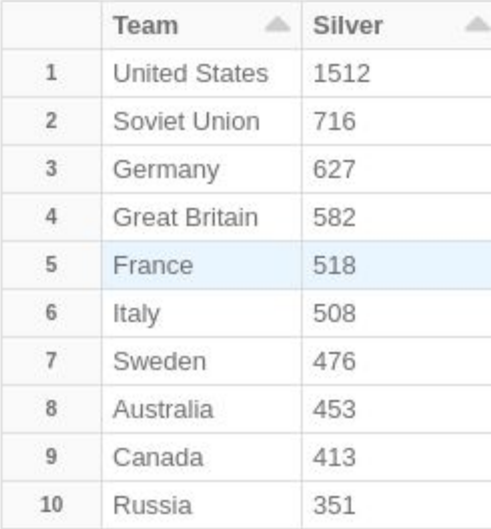
Thus, all the distributions follow the overall trend, that is, highest number of medalists from the age group of **23** and most athletes from the age group of **20-29** have most chances of winning.

Diving Deeper (Country)

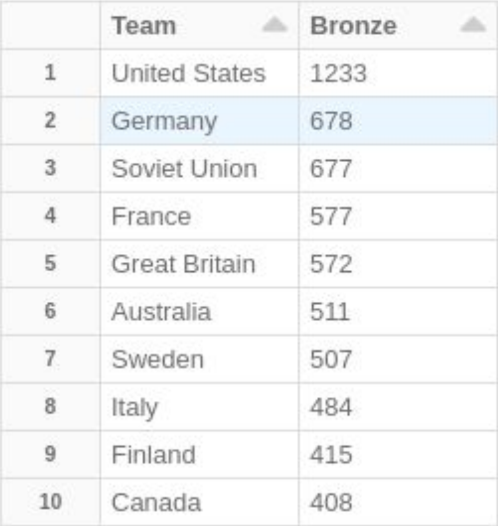
- Most number of medals (**5219**) goes to '**UNITED STATES**' team (**Gold - 2474**, **Silver - 1512**, **Bronze - 1233**). The average age of winning athletes is **24.6** years.



	Team	Gold
1	United States	2474
2	Soviet Union	1058
3	Germany	679
4	Italy	535
5	Great Britain	519
6	France	455
7	Sweden	451
8	Hungary	432
9	Canada	422
10	East Germany	369



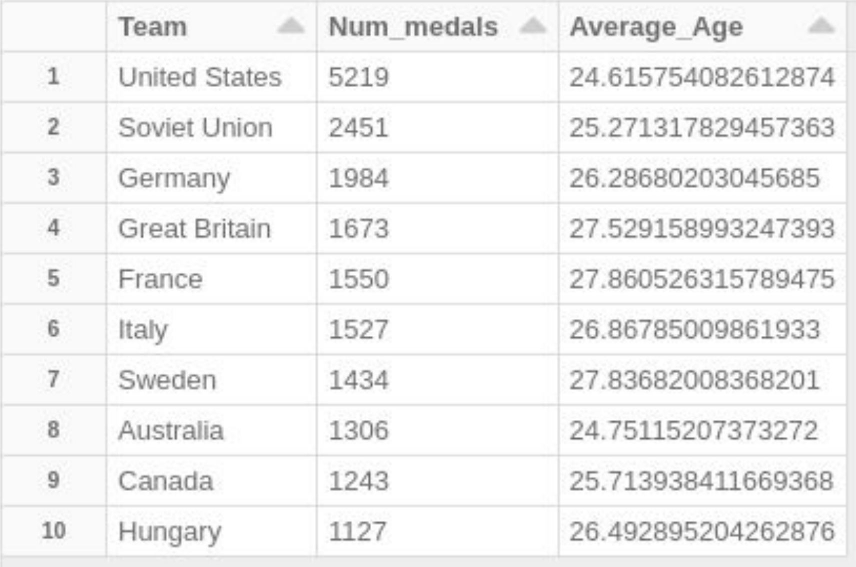
	Team	Silver
1	United States	1512
2	Soviet Union	716
3	Germany	627
4	Great Britain	582
5	France	518
6	Italy	508
7	Sweden	476
8	Australia	453
9	Canada	413
10	Russia	351



	Team	Bronze
1	United States	1233
2	Germany	678
3	Soviet Union	677
4	France	577
5	Great Britain	572
6	Australia	511
7	Sweden	507
8	Italy	484
9	Finland	415
10	Canada	408

Diving Deeper (Country) - cont.

- The average age of athletes from top 3 winning teams is: **24.6 years** (UNITED STATES), **25.3 years** (SOVIET UNION) and **26.3 years** (GERMANY).



	Team	Num_medals	Average_Age
1	United States	5219	24.615754082612874
2	Soviet Union	2451	25.271317829457363
3	Germany	1984	26.28680203045685
4	Great Britain	1673	27.529158993247393
5	France	1550	27.860526315789475
6	Italy	1527	26.86785009861933
7	Sweden	1434	27.83682008368201
8	Australia	1306	24.75115207373272
9	Canada	1243	25.713938411669368
10	Hungary	1127	26.492895204262876

Key points from the discovery…

- ➢ Most medals are won by athletes who are **23 years old.**
- ➢ Athletes in their twenties, that is, 20-29 years of age tend to win more medals than others.
- ➢ This relationship is also followed in the respective medal (GOLD, SILVER, BRONZE) groups.
- ➢ One more interesting relationship was between the athletes and their country**.** Developed countries tend to provide more attention to their sportspersons and hence the better performance.

Initial Hypothesis (Right or Wrong?)

My initial hypothesis was proven partially right and partially wrong.

- Instead of 24-27, the best age for an athlete is 23 or ranging from 20-29.
- There indeed is a relationship between the age of the athlete and his performance as athletes in the age group 20-29 tend to win more than the others.
- It’s still very early to say if this relationship applies to every athlete as we can see many older athletes winning medals. But we can say that this **GENERALLY** applies to everyone.

Addition to the Initial Hypothesis

- To win or to perform better, athletes need the support of their country.
- If the talented ones are nurtured from a young age, the performance will surely go up.
- Many governments still see sports as a hobby rather than a profession which needs to stop for the sake of young athletes and their future.

Summary and Recommendations

As it is always, youngsters are the key to success. It applies in Sports as well. The middle 20s is the age where any athlete is usually at the peak of his/her performance. Country of origin also has a say in any athletes’ performance. How the country has helped the athlete in his/her training is reflected in the athletes’ performance. Developed countries have understood that and are helping their athletes’ grow by putting resources in their training(making sure they train with some best trainers, have a healthy diet, get enough time to relax, learn the best techniques in the sports, etc.) and also helping them financially.

My recommendations are that talented athletes’ should be put into a training program by the country administration from a young age. They should learn from the best in the country. There should be events held to have a healthy competition within the academy. The country should motivate the young talents through campaigns/competitions to pursue a pro career if they have what it takes to be one.

THANK YOU!!