MILESTONE 4: PRESENTING YOUR FINDING

DATA ANALYSIS PROPOSAL FOR OLYMPICS DATASET

JULY / 2022

STEVEN LEONG

CONTENTS

Preparation Page

1. QUESTIONS TO ANSWER / HYPOTHESIS / APPROACH

QUESTIONS TO ANSWER

- HOW MANY PERUVIANS WON A GOLD MEDAL?
- HAS CRISTIANO RONALDO (CR7) WON ANY MEDALS?
- WHICH SPORTING EVENT GATHERED THE MOST ATHLETES?
- IN WHICH YEAR WERE THE MOST GOLD MEDALS AWARDED?
- IN THE 120 YEARS OF THE OLYMPIC GAMES, WHICH COUNTRY'S TEAM WON THE MOST MEDALS?
- ARE THERE MORE MEN OR WOMEN WHO HAVE WON A GOLD MEDAL?
- IF THE DATASET ONLY HAS EVENTS UP TO THE YEAR 2016, HOW DO YOU PREDICT WHICH MEDAL A NEW ATHLETE WILL WIN?

INITIAL HYPHOTHESIS

HOW MANY PERUVIANS WON A GOLD MEDAL?

YES, THERE ARE PERUVIANS WHO WON GOLD MEDALS

- **HAS EDWIN GONZALO VSQUEZ CAM WON ANY MEDALS?**

I THINK THAT HE HAS WON A GOLD MEDAL

- **WHICH SPORTING EVENT GATHERED THE MOST ATHLETES?**

WITHOUT A DOUBT, FOOTBALL IS THE SPORTING EVENT THAT GATHERS MORE ATHLETES

- **IN WHICH YEAR WERE THE MOST GOLD MEDALS AWARDED?**

I BELIEVE THAT MORE GOLD MEDALS WERE AWARDED BETWEEN THE YEARS 2006 TO 2010

- **IN THE 120 YEARS OF THE OLYMPIC GAMES, WHICH COUNTRY'S TEAM WON THE MOST MEDALS?**

FROM MY PERSPECTIVE, POLAND HAS MORE MEDALS

- **ARE THERE MORE MEN OR WOMEN WHO HAVE WON A GOLD MEDAL?**

THERE ARE MORE MALE ATHLETES, THEREFORE, I THINK THAT MEN WIN MORE MEDALS

- **IF THE DATASET ONLY HAS EVENTS UP TO THE YEAR 2016, HOW DO YOU PREDICT WHICH MEDAL A NEW ATHLETE WILL WIN?**

I HAVE NO IDEA WHAT WILL HAPPEN, BUT WE WILL FIND OUT USING MACHINE LEARNING LATER.

DATA ANALYSIS APPROACH

THE WORKING ENVIRONMENT WILL BE IN DATABRICKS PLATFORM TO CARRIED OUT THE FINDING RESULTS FROM THE OPEN DATASET.

IN THE FIRST INSTANCE, THE CSV FILE ATHLET_EVENTS.CSV WILL BE CARRIED OUT IN THE ANALYSIS EVENTS. THAT IS, THE COLUMN EXISTS IN BOTH FILES. THEN IT WILL BE NECESSARY TO REMOVE OR REPLACE THE NA VALUES FOR BETTER ANALYSIS.

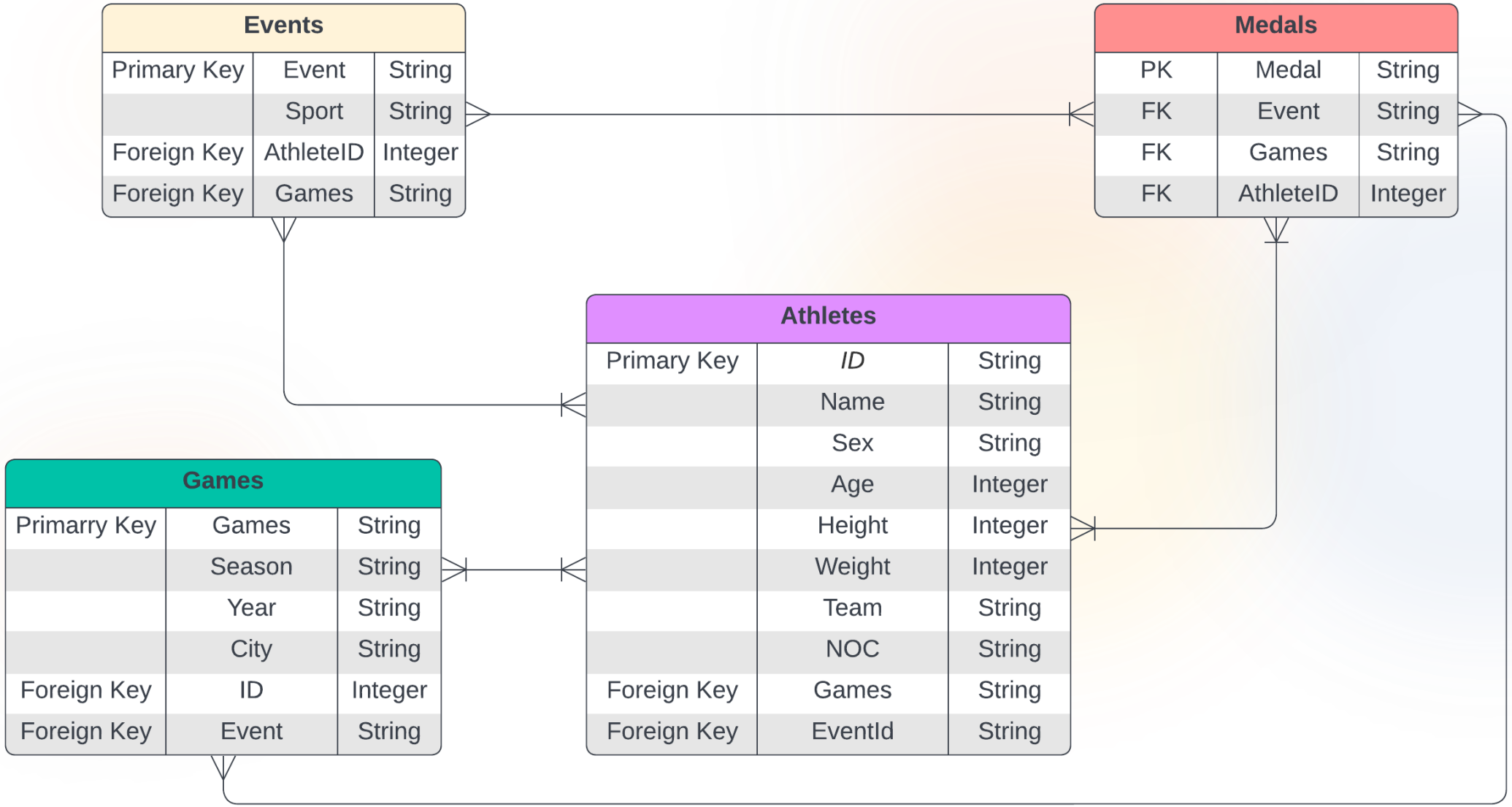
FINALLY, TO ANSWER THE QUESTIONS POSED, STATISTICAL INFERENCE AND GRAPHIC VISUALIZATION WILL BE USED TO DETERMINE IF THERE IS A RELATIONSHIP BETWEEN THE COLUMNS.

2. DISCUSS TECHNICAL CHALLENGES

- DATABRICKS WITH SQL WAS USED FOR ALL ANALYZES.
- DATABRICKS WAS THE PLATFORM WHICH INTEGRATE ALL MACHINE LEARNING AND DATA ANALYSIS ENVIRONMENT
- WE HAVE ESCALATE AND CLEAN DATA FOR THOSE NULL VALUES TO ‘NA’ IN ORDER TO HAVE DATA CONSISTENCY. ALL CALCULATIONS CARRIED OUT USING DATA WHICH WAS AVAILABLE TO PERFORM CALCULATION.

3. ERD

ERD Diagram (Athlete Sport Events)



4. INITIAL FINDINGS

CREATING OLYMPICS DATABASE FROM CSV DATASET

CSV FILE BASED OPEN DATASETS FOR ANALYSIS USING COMPUTE ENGINE AND IT WAS USING DELTA TABLE TO PERFORM DML MANIPULATION OF THE DATA:

```
CREATE TABLE AthleteEvents
USING DELTA
AS
SELECT * FROM athlete_events;
```

PERFORM DATA CLEANING PROCESS FOR AGE

```
UPDATE AthleteEvents SET Age='NA' WHERE Age IS NULL;
```

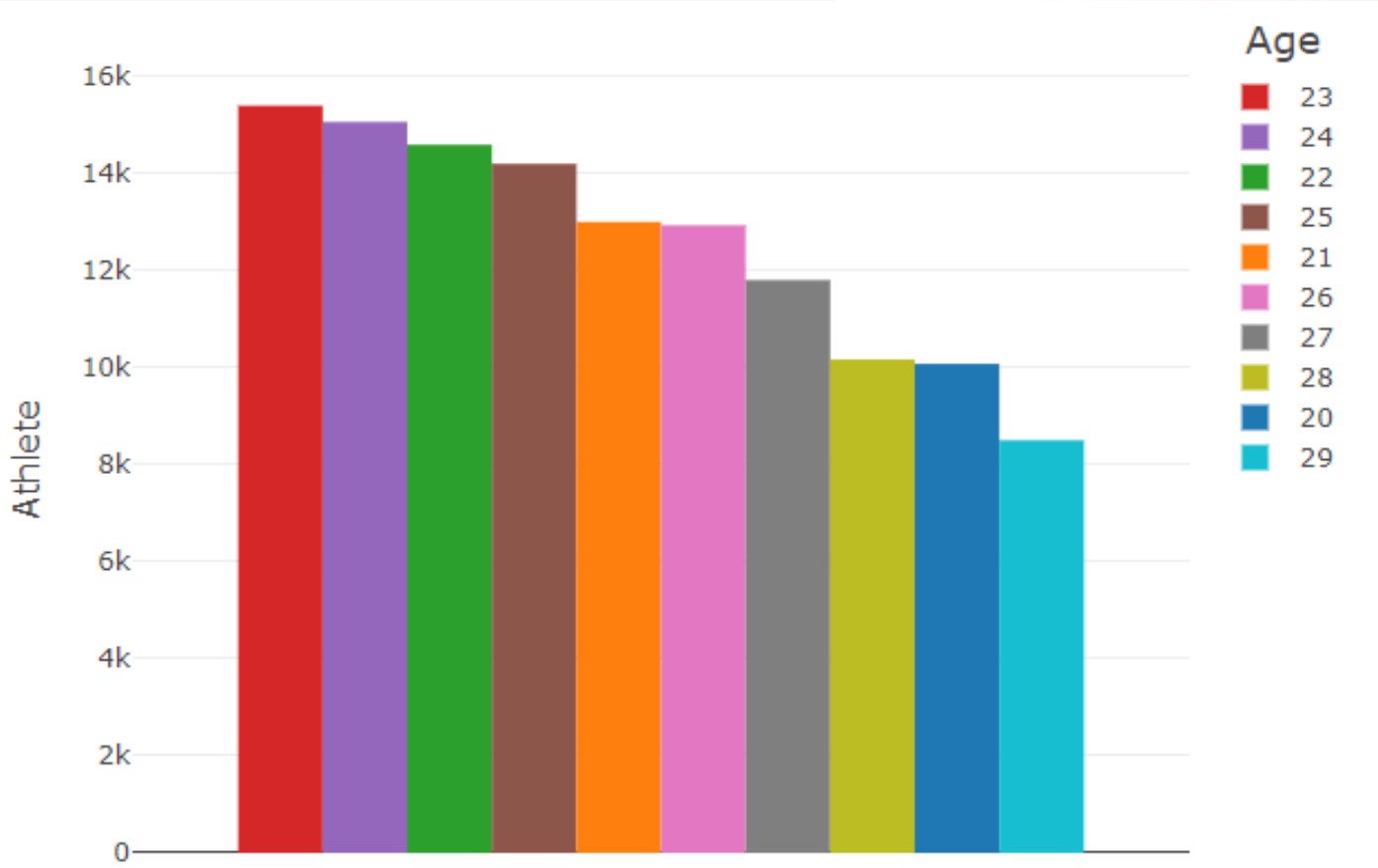
PREVIEW DATA FROM CSV TO DELTA TABLE

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

Year	Season	City	Sport	Event	Medal
1992	Summer	Barcelona	Basketball	Basketball Men's Basketball	NA
2012	Summer	London	Judo	Judo Men's Extra-Lightweight	NA
1920	Summer	Antwerpen	Football	Football Men's Football	NA
1900	Summer	Paris	Tug-Of-War	Tug-Of-War Men's Tug-Of-War	Gold
1988	Winter	Calgary	Speed Skating	Speed Skating Women's 500 metres	NA

DESCRIPTIVE ANALYSIS BY AGE

```
SELECT Age, COUNT(DISTINCT ID) As Athlete
FROM AthleteEvents WHERE Age NOT IN ('NA')
GROUP BY Age
ORDER BY COUNT(DISTINCT ID) DESC LIMIT 10;
```



- AGE

Age	Athlete
23	15400
24	15065
22	14595
25	14201
21	13003
26	12935
27	11806
28	10170
20	10079
29	8504

DESCRIPTIVE ANALYSIS BY ATHLETES

```
SELECT Year, City, Season, COUNT(DISTINCT ID) As Athletes, COUNT(DISTINCT Event) As Events, Count(Medal) As Medals
FROM AthleteEvents
WHERE Event IS NOT NULL AND Medal NOT IN ('NA') AND Team IS NOT NULL
GROUP BY Year, City, Season
ORDER BY Count(Medal) DESC;
```

Year	City	Season	Athletes	Events	Medals
2008	Beijing	Summer	1873	302	2048
2016	Rio de Janeiro	Summer	1855	306	2023
2000	Sydney	Summer	1839	300	2004
2004	Athina	Summer	1836	301	2001
2012	London	Summer	1772	302	1941
1996	Atlanta	Summer	1692	271	1842
1992	Barcelona	Summer	1544	257	1712
1988	Seoul	Summer	1396	237	1582
1984	Los Angeles	Summer	1316	221	1476
1980	Moskva	Summer	1254	203	1384
1976	Montreal	Summer	1186	198	1320
1920	Antwerpen	Summer	1074	156	1308
1972	Munich	Summer	1070	193	1215
1968	Mexico City	Summer	921	172	1057
1964	Tokyo	Summer	920	163	1029

5. DEEPER ANALYSIS

How Many Peruvians won a gold medal

```
SELECT * FROM AthleteEvents WHERE NOC='PER' AND Medal='Gold';
```

ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
125468	Edwin Gonzalo Vsquez Cam	M	25	NA	NA	Peru	PER	1948 Summer	1948	Summer	London	Shooting	Shooting Men's Free Pistol, 50 metres	Gold

Has Edwin Gonzola Vsquez Cam won any medals?

```
SELECT * FROM AthleteEvents WHERE NOC='PER' AND Name LIKE 'Edwin Gonzalo%';
```

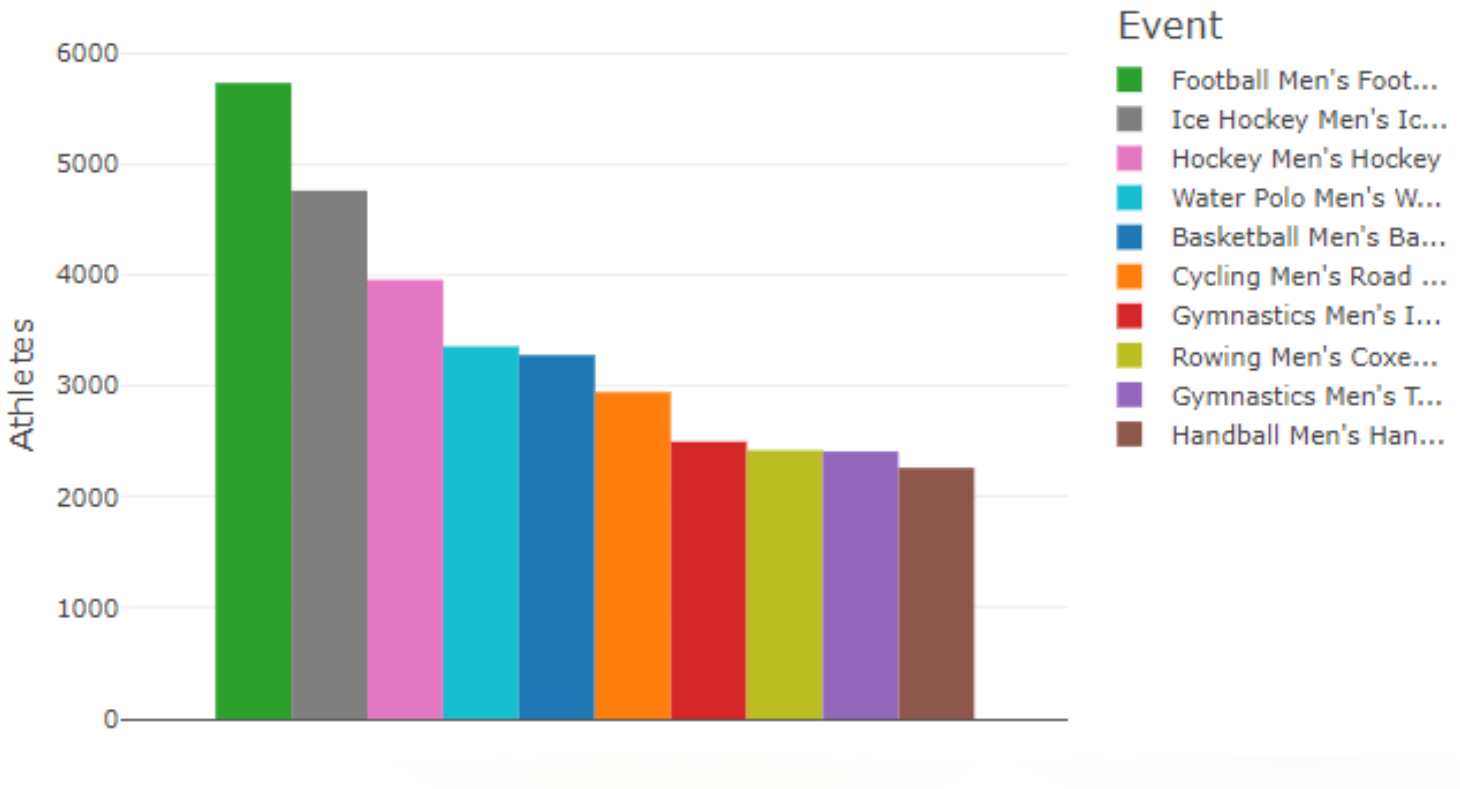
ID	Name	Sex	Age	Height	Weight	Team	NOC	Games	Year	Season	City	Sport	Event	Medal
125468	Edwin Gonzalo Vsquez Cam	M	25	NA	NA	Peru	PER	1948 Summer	1948	Summer	London	Shooting	Shooting Men's Free Pistol, 50 metres	Gold

DEEPER ANALYSIS

Which sport event gathered the most athletes from events?

```
SELECT Event, COUNT(Id) as Athletes
FROM AthleteEvents
GROUP BY Event
ORDER BY Athletes DESC LIMIT 10;
```

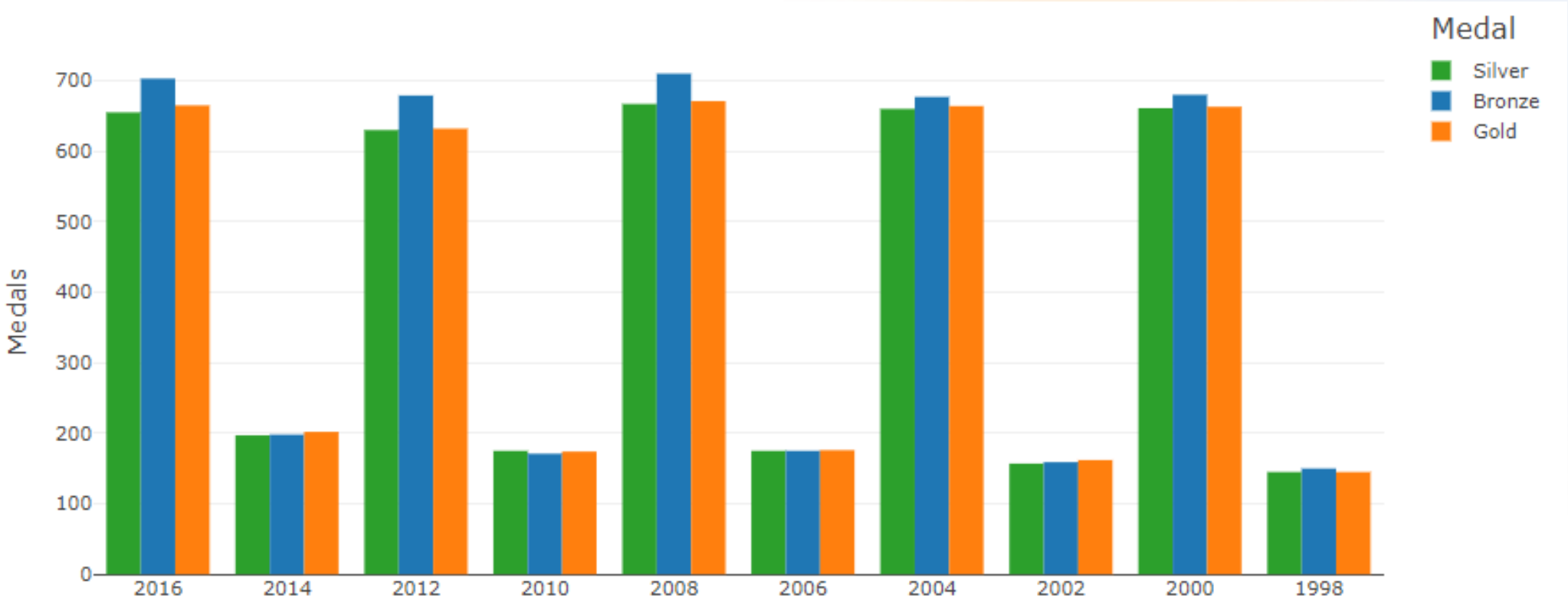
Event	Athletes
Football Men's Football	5733
Ice Hockey Men's Ice Hockey	4762
Hockey Men's Hockey	3958
Water Polo Men's Water Polo	3358
Basketball Men's Basketball	3280
Cycling Men's Road Race, Individual	2947
Gymnastics Men's Individual All-Around	2500
Rowing Men's Coxed Eights	2423
Gymnastics Men's Team All-Around	2411
Handball Men's Handball	2264



DEEPER ANALYSIS

In which year were the most award to gold medal

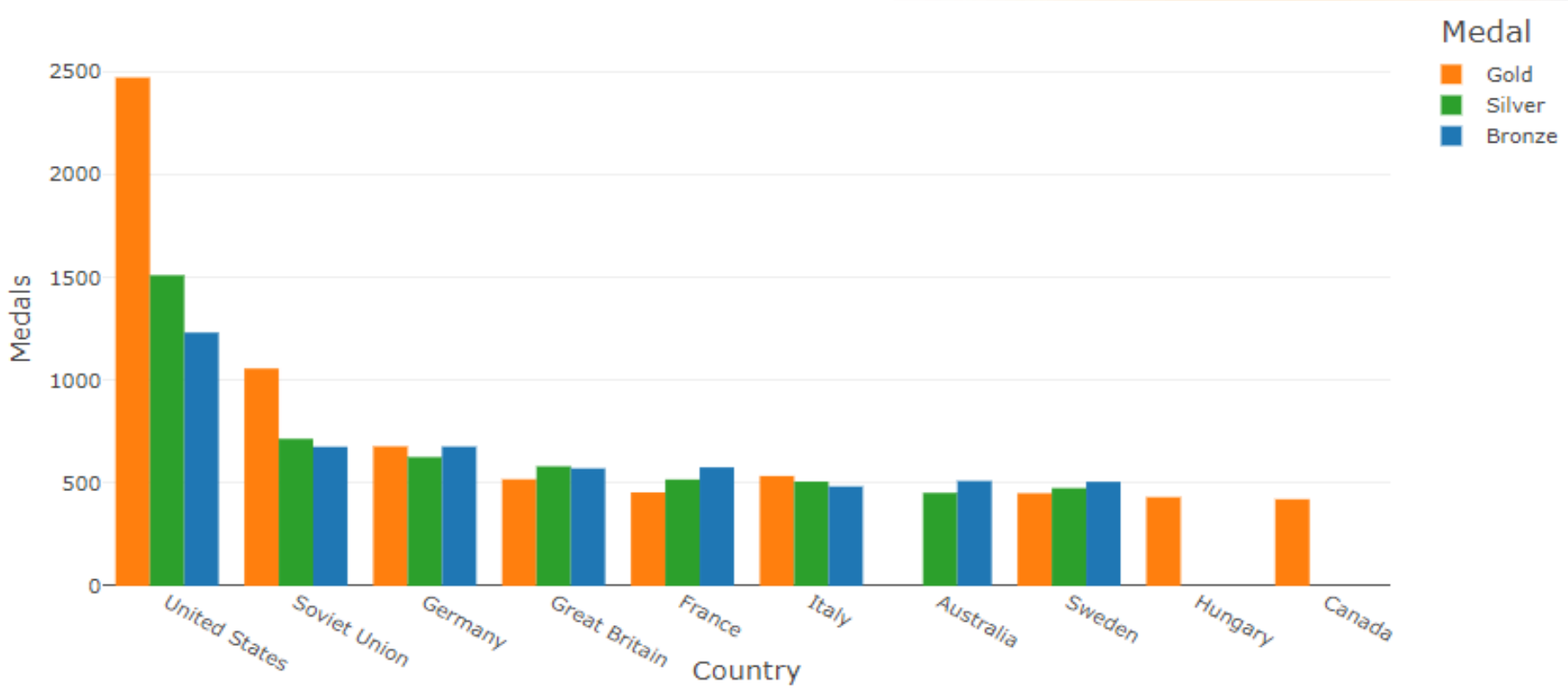
```sql
SELECT Year, Medal, COUNT(Id) As Medals
FROM AthleteEvents
WHERE Medal NOT IN ('NA')
GROUP BY Year, Medal
ORDER BY Year DESC LIMIT 30;
```



DEEPER ANALYSIS

In the 120 years of Olympic Games, which country team won the most medals

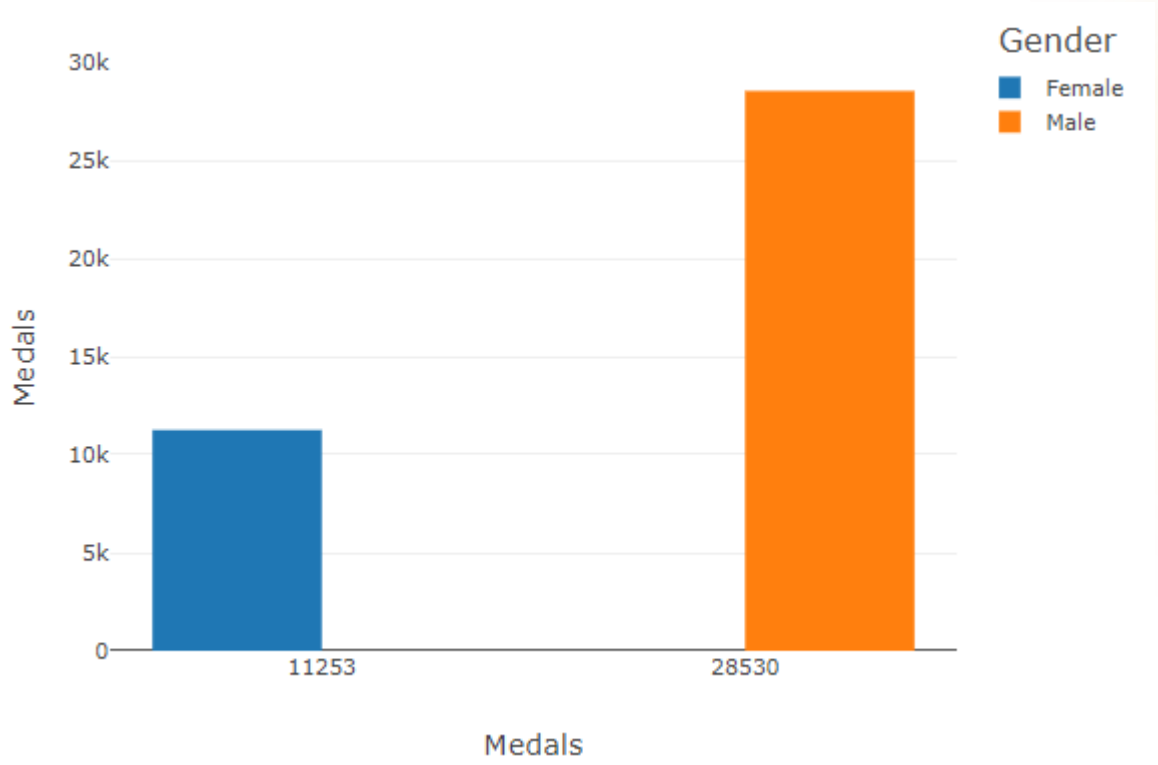
```sql
SELECT Team As Country, Medal, COUNT(Id) As Medals
FROM AthleteEvents
WHERE Medal NOT IN ('NA')
GROUP BY Country, Medal
ORDER BY Medals DESC LIMIT 25;
```



DEEPER ANALYSIS

Comparison gender who won a gold medal?

```
SELECT CASE WHEN Sex='M' THEN 'Male' ELSE 'Female' END AS Gender, COUNT(Medal) As Medals
FROM AthleteEvents
WHERE Medal NOT IN ('NA')
GROUP BY Sex;
```



6. HYPHOTHESIS RESULTS

HOW MANY PERUVIANS WON A GOLD MEDAL?
ONLY ONE PERUVIAN HAS WON A GOLD MEDAL

- **HAS EDWIN GONZALO VSQUEZ CAM WON ANY MEDALS?**
YES, HE HAS WON ONE GOLD MEDAL AT THE OLYMPIC GAMES

- **WHICH SPORTING EVENT GATHERED THE MOST ATHLETES?**
FOOTBALL HAS MORE ATHLETES

- **IN WHICH YEAR WERE THE MOST GOLD MEDALS AWARDED?**
IN 2008, MORE GOLD MEDALS WERE AWARDED

- **IN THE 120 YEARS OF THE OLYMPIC GAMES, WHICH COUNTRY'S TEAM WON THE MOST MEDALS?**
USA HAS MORE MEDALS

- **ARE THERE MORE MEN OR WOMEN WHO HAVE WON A GOLD MEDAL?**
LIKE I SAID, THERE ARE MORE MEN WITH GOLD MEDALS

- **IF THE DATASET ONLY HAS EVENTS UP TO THE YEAR 2016, HOW DO YOU PREDICT WHICH MEDAL A NEW ATHLETE WILL WIN?**
A MACHINE LEARNING MODEL HAS BEEN CREATED TO PREDICT NEW DATA FROM TRAINING WITH EXISTING DATA.

7. RECOMMENDATIONS

POSITIVE, LINEAR AND STRONG CORRELATION BETWEEN HEIGHT AND WEIGHT. THE WEIGHT VARIABLE IS A SIGNIFICANT PREDICTOR OF THE HEIGHT VARIABLE. THE OBSERVED STUDIES LACK EXPERIMENTAL DESIGN PRINCIPLES. THEREFORE, THE FINDINGS ON THE DATA CANNOT INFER CAUSALITY. OBSERVATIONAL STUDIES CAN ONLY INFER CORRELATION