

120 Years of Olympics Dataset Analysis

Learn SQL Basics Capstone

Milestone 4

# Contents of this Presentation

- Overview of Dataset and Potential Clients
- Questions I want to answer
- Initial Hypotheses
- Approach taken
- Entity Relationship Diagram
- Descriptive Statistics and Initial Exploration
- Dive Deeper, Go Broader Analysis
- Results of Hypotheses
- Further Plan of Study



# Overview of Dataset and Potential Clients


## Overview:

- Historic Dataset on Modern Olympic Games
- Covers all tournaments from Athens 1896 to Rio 2016

## Reason for Selection:

- Interest in Olympic games and digging patterns in different sports.

## Target Audience:

- SportsStats, a sports analysis firm partnering with news channels, sports experts.
  - Purpose: Developing a news story/Discovering key insights.
- 

# Questions I want to answer....

1. Which countries have won the most number of medals? How has this changed over the years? Does being a host country increase the chance of winning a medal?

This will help us understand the well performing nations in the olympics.

2. Why is there an alternating high and low participation of players?

This will help uncover the different types of olympic games.

3. Do new games get introduced in the olympics? If yes then which games were recently introduced?
4. Which sports correspond to greater height and weight of the players? Does this match our intuition? Do they win more medals than others?

This can be useful info for news articles.



# Initial Hypotheses....

1. I expect The United States, China and Japan to have won the highest number of medals.

This is by looking at the medal Tally in the Tokyo Olympics 2020 this year.

2. This alternating pattern is due to a different type of Olympic Game played.

I think so because all other factors look same.

3. I expect that new games do get introduced in the Olympics.

This is because newer sports keep getting developed.

4. I expect games like basketball to have greater height of players, because

This matches with our intuition.



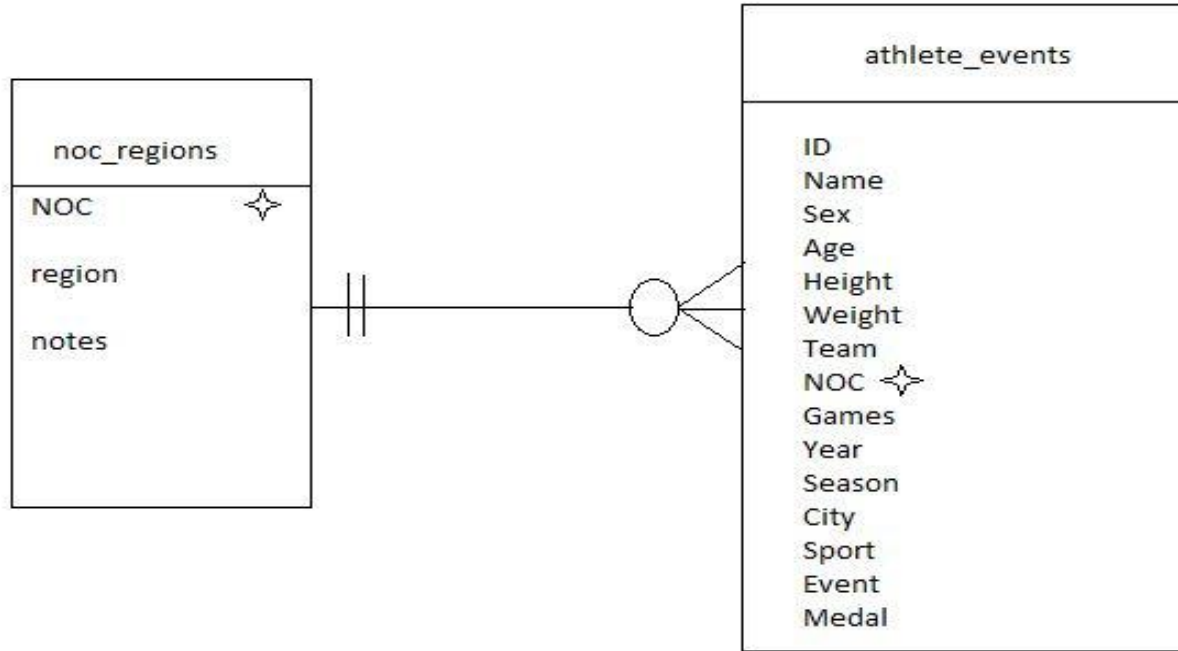
# Approach Taken....

What quantities will I look at?

1. i) Number of rows grouped by country.  
ii) Ratio of number of players won to number of players participating when the country is a host and when it is not.
2. i) A table and plot of number of distinct players grouped by year.
3. i) Minimum date grouped by sport to get the date introduced.
4. i) Average height and weight for basketball and non-basketball players.



# Entity Relationship Diagram (ERD)....

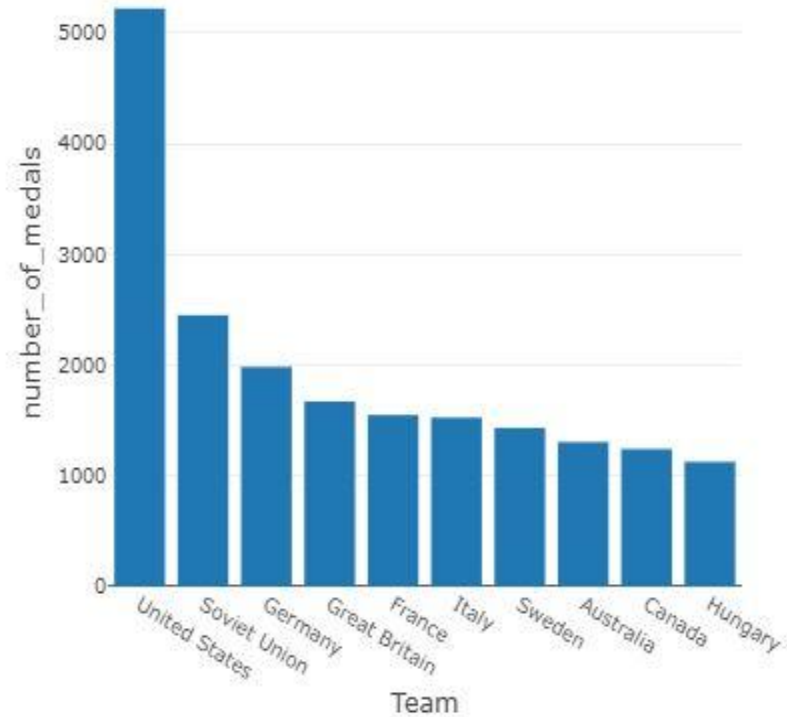


# Descriptive Statistics and Initial Exploration....

- Highest number of medals by

country:

The United States, Soviet Union, Germany, Great Britain, France have won the most medals.





# Descriptive Statistics Continued...

Performance of The United States in the summer olympics over the years..



# Descriptive Statistics Continued...

- Stats for Basketball Players:

	avg_height ▲	std_height ▲	avg_weight ▲	std_weight ▲	count ▲
1	190.63942489595158	11.384468415007088	85.03405221339388	14.503529228033813	2643

- Stats for Non-Basketball Players:

	avg_height ▲	std_height ▲	avg_weight ▲	std_weight ▲	count ▲
1	175.95498731818418	10.075671969151784	71.5616232724261	14.356605382812507	96595

Thus, Basketball players are significantly taller than Non-Basketball players.

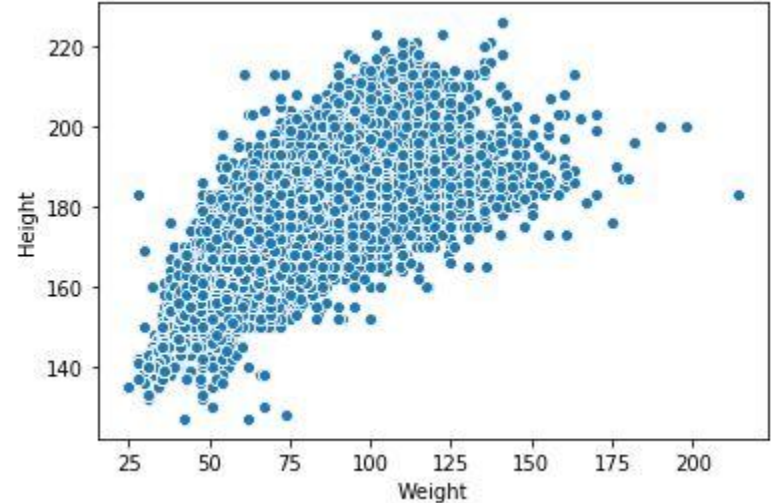


# Dive Deeper, Go Broader Analysis...

- Relationship between Height and Weight.....

A correlation coefficient of 0.79 between Height and Weight is observed...

- Similarly, no significant correlation is observed between age and height and Age and weight.



# Dive Deeper, Go Broader Analysis Continued..

- Relationship between Height and Country for both genders...

Men from Montenegro and Women from Croatia are the tallest on average in this dataset with average heights of 190 cm and 174 cm respectively:

```
SELECT Team, Sex, COUNT(*) AS num_players, AVG(height) AS avg_height, AVG(weight) AS avg_weight, AVG(age) AS avg_age
FROM
  (SELECT ID, Team, Sex, AVG(Height) AS height, AVG(Weight) AS weight, AVG(Age) AS age
   FROM aec
   WHERE (Height IS NOT NULL) AND (WEIGHT IS NOT NULL) AND (Age IS NOT NULL)
   GROUP BY ID, Team, Sex)
GROUP BY Team, Sex
HAVING num_players>100 AND Sex = 'M'
ORDER BY avg_height DESC
```

▶ (3) Spark Jobs

	Team	Sex	num_players	avg_height	avg_weight	avg_age
1	Serbia and Montenegro	M	164	190.90243902439025	87.73780487804878	25.097459349593493
2	Serbia	M	153	190.16993464052288	88.09150326797386	26.098848428260197
3	Croatia	M	287	189.58536585365854	90.2404181184669	25.972714451634314
4	Lithuania	M	199	188.67336683417085	86.19597989949749	25.70646486400255
5	Iceland	M	148	185.1891891891892	85.32432432432432	25.3381542256654222
6	Netherlands	M	1102	184.5880217785844	80.33212341197822	25.754039510541052
7	Czech Republic	M	467	183.66595289079228	82.50535331905782	26.66551455607018

```
SELECT Team, Sex, COUNT(*) AS num_players, AVG(height) AS avg_height, AVG(weight) AS avg_weight, AVG(age) AS avg_age
FROM
  (SELECT ID, Team, Sex, AVG(Height) AS height, AVG(Weight) AS weight, AVG(Age) AS age
   FROM aec
   WHERE (Height IS NOT NULL) AND (WEIGHT IS NOT NULL) AND (Age IS NOT NULL)
   GROUP BY ID, Team, Sex)
GROUP BY Team, Sex
HAVING num_players>100 AND Sex = 'F'
ORDER BY avg_height DESC
```

▶ (3) Spark Jobs

	Team	Sex	num_players	avg_height	avg_weight	avg_age
1	Croatia	F	110	174.55454545454546	65.71818181818182	24.000303030303034
2	Yugoslavia	F	129	173.4108527131783	65	22.82700761770529
3	Netherlands	F	669	172.74140508221225	64.57249626307922	24.680710657055947
4	Czech Republic	F	244	172.7377049180328	63.131147540983605	24.308762538222513

# Dive Deeper Go Broader Continued....

- Recently introduced sports include Rugby Sevens, Triathlon, Trampolining, Taekwondo...

```
SELECT Sport, MIN(Year) AS year_introduced
FROM
  (SELECT ID, Name, Sex, Age, Height, Weight, Team, NOC, Games, Year, Season, City, Sport, Event, Medal, COUNT(*) AS count
   FROM aec
   GROUP BY ID, Name, Sex, Age, Height, Weight, Team, NOC, Games, Year, Season, City, Sport, Event, Medal
   HAVING count=1)
GROUP BY Sport
ORDER BY year_introduced DESC
```

	Sport ▲	year_introduced ▲
1	Rugby Sevens	2016
2	Triathlon	2000
3	Trampolining	2000
4	Taekwondo	2000
5	Snowboarding	1998
6	Softball	1996
7	Beach Volleyball	1996

# Dive Deeper Go Broader Continued...

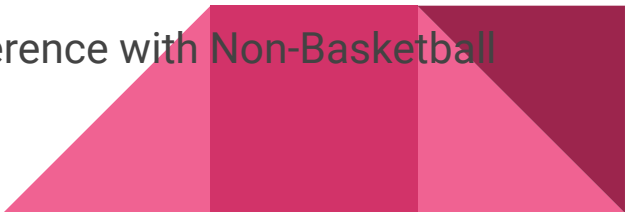
- 2 new metrics were created to analyze the effect of being a host country on medal-winning:
  1. Ratio of medals won by players of a particular country to total medals won that year.
  2. Ratio of medals won by players of a particular country to total participants of that country.

These 2 metrics were calculated for 2 cases: When the country is a host country vs when it is not for different countries.

- China seems to improve its performance the most when it is a host country.



# Results of Hypotheses...

1. Yes, The United States has won the most number of medals so far. However, China and Japan are not in the top 5 list which was expected. Also, being a host country is advantageous to China the most.
  2. The alternating high and low number of players is due to Summer and Winter games being played.
  3. Yes, new games do get introduced in the Olympics and Rugby Sevens is the latest introduced player.
  - 4.. Yes, Basket Players have a statistically significant height difference with Non-Basketball players..
- 

# Further Plan of Study

Some possible options for further analysis...

1. If possible, collect other data about the players fitness and merge these 2 datasets. Then try to find insights.
2. Predict whether or not a player will win a medal in a particular game given all these features.
  - My client now has plenty of descriptive statistics and insights to give to nws channels or fitness trainers.





---

**THANK  
YOU**

---