**1. Summary of Descriptive Statistics**

To gain a comprehensive understanding of the Olympic dataset, several descriptive statistics were examined. These initial analyses are crucial for shaping the direction of our research and ensuring the validity of our findings.

* **Counts of Categorical Data (Sport, Season, Medal, NOC):** We began by counting the occurrences of unique values in key categorical columns.
  + **Why:** This helps us understand the scope and distribution of the data. For example, knowing the number of unique sports tells us the breadth of athletic disciplines covered, while counting medals per National Olympic Committee (NOC) gives a preliminary look at country-level performance. The distribution of medal types (Gold, Silver, Bronze, and NA) is fundamental to nearly all performance-related questions.
* **Range of Numerical Data (Year):** We identified the minimum and maximum years in the dataset.
  + **Why:** This establishes the historical timeframe of our data (1896-2016), which is essential for analyzing trends and defining different Olympic eras as proposed in our research questions.
* **Measures of Central Tendency (Average Age, Height, Weight):** We calculated the average age, height, and weight of all athletes.
  + **Why:** These averages provide a baseline "typical" athlete profile. This is the first step in exploring our hypothesis about how the physical characteristics of athletes have evolved over time and across different sports.
* **Frequency Distributions (Top 10 NOCs by Medal Count):** We generated a ranked list of the top 10 NOCs based on the total number of medals won.
  + **Why:** This directly addresses our first research question about identifying dominant nations. It provides a clear, high-level overview of which countries have historically been the most successful in the Olympics.

**Data Cleaning and Preparation**

Before proceeding with in-depth analysis, the dataset was rigorously cleaned to ensure data quality and accuracy. The following steps were performed:

* **Handling Missing Values:** Null values in the numerical columns Age, Height, and Weight were imputed using the overall average for each respective column. This ensures that these records can be included in numerical analyses without skewing the results.
* **Removing Duplicates:** A query was run to identify and remove fully duplicate rows. This process eliminated 1,385 redundant records, preventing overcounting in our analysis.
* **Data Validation:** Checks were performed to ensure there were no null values in critical categorical columns (e.g., Name, NOC, Sport) and no invalid data points (e.g., negative values for age or weight).
* **Final Record Count:** After cleaning, the dataset contains **269,731** records, reduced from the initial 271,116. This cleaned dataset forms the basis for all subsequent analysis.

**SQL Queries applied on the data**

* --Playing with Data
* --Clean the data--deal with NULLS-replace with average for numeric values like age height and weight
* SELECT \*
* FROM athlete\_events
* --NULL Values
* SELECT
* COUNT (\*) AS Total,
* COUNT(\*)-COUNT(Height) AS Null\_Height,
* COUNT(\*)-COUNT(Age) AS Null\_Age,
* COUNT(\*)-COUNT(Weight) AS Null\_Weight
* FROM athlete\_events
* --WHERE Age IS NULL
* --Sports in which USA won medal (58 out of 66)
* SELECT DISTINCT Sport
* FROM athlete\_events
* WHERE Medal IS NOT NULL AND NOC ='USA'
* --NOC 230 but total team count is 1184
* SELECT COUNT (DISTINCT Team) AS Total
* FROM athlete\_events
* --Team names.....some weird team names like alibabaXI,Rambo won medal? others are understandable like Japan 1 and Japan 2...where two or more teams more Japan participate in same event
* SELECT Team,
* COUNT(\*) AS Total
* FROM athlete\_events
* WHERE Medal IS NOT NULL
* GROUP BY Team
* ORDER BY Total ASC
* --Finding out where Team Rambo won medal
* SELECT Sport,
* YEAR,
* Event
* FROM athlete\_events
* WHERE TEAM ='RAMBO'
* --Starting with Cleaning the data
* --1.Identify Null Values for Age/Weight/height and replacing them with average values as counted before (Age 25.055,Height-175.371,Weight-70.688) for Numeric values except year offcourse
* SELECT
* AVG(Age) AS Average\_Age,
* AVG(Height) AS Average\_Height,
* AVG(Weight) AS Average\_Weight
* FROM athlete\_events
* WHERE Age IS NOT NULL AND Height IS NOT NULL AND Weight IS NOT NULL
* UPDATE athlete\_events
* SET Age = 25.0555089370165
* WHERE Age is NULL;
* UPDATE athlete\_events
* SET Height = 175.371949651978
* WHERE Height is NULL;
* UPDATE athlete\_events
* SET Weight = 70.6883370116169
* WHERE Weight is NULL;
* --Null values for non-numeric Data except Medals as NA means not 1st,2nd or 3rd place finish in event-------------0 null values..
* SELECT
* COUNT (\*) AS Total,
* COUNT(\*)-COUNT(Name) AS Null\_Name,
* COUNT(\*)-COUNT(Sex) AS Null\_Sex,
* COUNT(\*)-COUNT(Team) AS Null\_Team,
* COUNT(\*)-COUNT(NOC) AS Null\_NOC,
* COUNT(\*)-COUNT(Games) AS Null\_Games,
* COUNT(\*)-COUNT(Year) AS Null\_Year,
* COUNT(\*)-COUNT(Season) AS Null\_Season,
* COUNT(\*)-COUNT(City) AS Null\_City,
* COUNT(\*)-COUNT(Sport) AS Null\_Sport,
* COUNT(\*)-COUNT(Event) AS Null\_Event
* FROM athlete\_events
* --2. Identifying Duplicate Data
* SELECT Id,
* Name,
* Sex,
* Age,
* Height,
* Weight,
* Team,
* NOC,
* Games,
* Year,
* Season,
* City,
* Sport,
* Event,
* Medal,
* COUNT(\*) AS Duplicates
* FROM athlete\_events
* GROUP BY Id,
* Name,
* Sex,
* Age,
* Height,
* Weight,
* Team,
* NOC,
* Games,
* Year,
* Season,
* City,
* Sport,
* Event,
* Medal
* HAVING COUNT(\*) > 1
* --Deleting Duplicate data using CTE -1385 Rows affected
* --Checking
* WITH CTE\_Duplicates AS (
* SELECT
* Id, Name, Sex, Age, Height, Weight, Team, NOC, Games, Year, Season, City, Sport, Event, Medal,
* ROW\_NUMBER() OVER (
* PARTITION BY Id, Name, Sex, Age, Height, Weight, Team, NOC, Games, Year, Season, City, Sport, Event, Medal
* ORDER BY Id
* ) as rn
* FROM athlete\_events
* )
* SELECT \* -- See which rows would be marked for deletion
* FROM CTE\_Duplicates
* WHERE rn > 1;
* --Deleting
* WITH CTE\_Duplicates AS (
* SELECT
* Id,
* Name,
* Sex,
* Age,
* Height,
* Weight,
* Team,
* NOC,
* Games,
* Year,
* Season,
* City,
* Sport,
* Event,
* Medal,
* ROW\_NUMBER() OVER (
* PARTITION BY
* Id, Name, Sex, Age, Height, Weight, Team, NOC, Games, Year, Season, City, Sport, Event, Medal
* ORDER BY Id
* ) as rn
* FROM
* athlete\_events
* )
* DELETE FROM CTE\_Duplicates
* WHERE rn > 1;
* -- Get the total number of records Raw Data=2,71,116 and now it is 2,69,731
* SELECT COUNT(\*) AS total\_records
* FROM athlete\_events
* --3.Fixing Data Type Inconsistencies like if a column is storing numbers as text, or dates in an incorrect format.
* Select \*
* From athlete\_events
* --None Visible
* --4. Standardizing Data Formats like Trim Whitespace: Remove leading/trailing spaces/Case issues/Date formats/Synonyms/Spelling Mistake
* Select \*
* From athlete\_events
* --None Visible
* --5.Correcting Invalid Data (Outliers, Incorrect Values) like Update Incorrect Values: Based on business rules or external data-----No entries with negative value
* SELECT Age,
* Height,
* Weight
* FROM athlete\_events
* WHERE Age <= 0 OR Height <=0 OR Weight <=0
* --Database is exported to Excel as Cleaned.xlsx

**2. Key Discoveries and New Ideas**

Our initial exploration has already yielded some interesting insights and sparked new questions.

1. **The Scale of "No Medal":** A significant finding was the sheer number of participants that did not result in a medal (231,333 "NA" entries in the original data).
   * Aha! Moment: While expected, the magnitude of this number emphasizes that our analysis should not only focus on medal winners. There is a rich story to be told about participation itself.
   * New Idea: We could investigate the "journeyman" athlete phenomenon—athletes who compete in multiple Olympics without winning a medal. What are their characteristics? Which NOCs have the highest number of non-medal-winning participants?
2. **Initial NOC Dominance:** The "Top 10 NOCs by Medal Count" clearly shows the long-standing dominance of a few nations, like the USA.
   * Aha! Moment: This confirms a widely held belief but also serves as a starting point for more nuanced questions. Is this dominance consistent across all sports, or is it concentrated in a few high-medal-count disciplines?
   * New Idea: We should analyze the medal count *per sport* for these top NOCs. This could reveal strategic specialization and challenge the idea of overall dominance. For example, does a country like Norway, not in the overall top 10, dominate Winter sports?
3. **Anomalies in Team Data:**
   * Aha! Moment: The number of unique Team names (1,184) is far greater than the number of unique NOCs (230). Investigation revealed unusual team names (e.g., 'Rambo', 'AlibabaXI') alongside standard national team variations (e.g., 'Japan-1', 'Japan-2'). This suggests that the Team column represents specific squads or historical team names that may not directly map to a single NOC for all entries.
   * New Idea: This discovery highlights the importance of using the NOC column for consistent country-level analysis. A deeper dive could be done to understand the context of these unique team names and whether they represent specific clubs, invitational teams, or historical naming conventions.

**3. Status of Initial Hypotheses**

**Hypothesis 1 (Partially Supported):** *Countries with larger populations and higher GDPs will generally accumulate more Olympic medals overall, but smaller, specialized nations may dominate specific niche sports.*

**Status:** The initial finding that a few large nations dominate the overall medal count supports the first part of this hypothesis.

**Next Steps:** To fully test this, we need to join our Olympic data with external datasets on the country’s population and GDP. For the second part, we will need to define "niche sports" and analyze the medal distribution within them to see if smaller nations indeed come out on top.

**Hypothesis 2 (Not Yet Tested):** *Over time, the average age of Olympic medalists in physically demanding sports has decreased, while in technically demanding or endurance sports, it may have increased.*

**Status:** Our calculation of the overall average age is just the first step. We have not yet broken this down by sport or year.

**Next Steps:** We will need to categorize sports (e.g., "physically demanding," "technical"). Then, for each category, we will plot the average age of medalists against the Olympic year to identify trends.

**Hypothesis 3 (Not Yet Tested):** *There will be a clear shift in the distribution of medals from traditional Western powers to emerging Asian and African nations in certain sports over the last few decades.*

**Status:** This hypothesis requires a time-series analysis that we have not yet performed.

**Next Steps:** We will define "eras" (e.g., pre-1980 vs. post-1980) and compare the medal distribution for specific sports between these periods. We will focus on the medal shares of different continents to see if a shift is evident.

**4. Additional Questions**

Our initial analysis has prompted the following new questions:

* How does the performance of female athletes compare to that of male athletes within the same sports and for the same NOCs over time?
* Are there "golden eras" for specific countries in specific sports (e.g., the Italian fencing teams of the 1980s)?
* What is the correlation between an athlete's age, height, and weight and their likelihood of winning a medal in specific events?