**Performing more Descriptive Stats**

**--------------------------------------------------------------------------------**

**-- Section 1: Overall Summary Counts**

**-- This section provides a high-level overview of the dataset's scope.**

**--------------------------------------------------------------------------------**

SELECT

COUNT(\*) AS TotalRecords,

COUNT(DISTINCT ID) AS TotalUniqueAthletes,

COUNT(DISTINCT Games) AS TotalOlympicGames,

COUNT(DISTINCT NOC) AS TotalCountries,

COUNT(DISTINCT Sport) AS TotalSports,

COUNT(DISTINCT Event) AS TotalEvents

FROM

athlete\_events;

**--------------------------------------------------------------------------------**

**-- Section 2: Categorical Data Distributions**

**-- This section looks at the distribution of key categorical variables.**

**--------------------------------------------------------------------------------**

**-- Distribution of Medals**

SELECT

Medal,

COUNT(\*) AS Count

FROM

athlete\_events

GROUP BY

Medal

ORDER BY

Count DESC;

**-- Distribution of Season**

SELECT

Season,

COUNT(\*) AS Count

FROM

athlete\_events

GROUP BY

Season;

**-- Distribution of Gender**

SELECT

Sex,

COUNT(\*) AS Count

FROM

athlete\_events

GROUP BY

Sex;

**--------------------------------------------------------------------------------**

**-- Section 3: Numerical Data Profile (Age, Height, Weight)**

**-- This section calculates key statistics for the athletes' physical attributes.**

**--------------------------------------------------------------------------------**

SELECT DISTINCT

'Age' AS Attribute,

AVG(Age) OVER () AS Average,

STDEV(Age) OVER () AS StandardDeviation,

MIN(Age) OVER () AS Minimum,

MAX(Age) OVER () AS Maximum,

PERCENTILE\_CONT(0.5) WITHIN GROUP (ORDER BY Age) OVER () AS Median

FROM athlete\_events

UNION ALL

SELECT DISTINCT

'Height' AS Attribute,

AVG(Height) OVER () AS Average,

STDEV(Height) OVER () AS StandardDeviation,

MIN(Height) OVER () AS Minimum,

MAX(Height) OVER () AS Maximum,

PERCENTILE\_CONT(0.5) WITHIN GROUP (ORDER BY Height) OVER () AS Median

FROM athlete\_events

UNION ALL

SELECT DISTINCT

'Weight' AS Attribute,

AVG(Weight) OVER () AS Average,

STDEV(Weight) OVER () AS StandardDeviation,

MIN(Weight) OVER () AS Minimum,

MAX(Weight) OVER () AS Maximum,

PERCENTILE\_CONT(0.5) WITHIN GROUP (ORDER BY Weight) OVER () AS Median

FROM athlete\_events;

**--------------------------------------------------------------------------------**

**-- Section 4: Top Performers**

**-- This section identifies the most successful countries and athletes.**

**--------------------------------------------------------------------------------**

-- Top 10 Countries by Total Medals

SELECT TOP 10

NOC,

COUNT(Medal) AS TotalMedals

FROM

athlete\_events

WHERE

Medal IS NOT NULL

GROUP BY

NOC

ORDER BY

TotalMedals DESC;

-- Top 10 Athletes by Total Medals

SELECT TOP 10

Name,

Team,

COUNT(Medal) AS TotalMedals

FROM

athlete\_events

WHERE

Medal IS NOT NULL

GROUP BY

Name, Team

ORDER BY

TotalMedals DESC;

**--------------------------------------------------------------------------------**

**-- Section 5: Participation Trends Over Time**

**-- This section analyzes how participation has changed across the years.**

**--------------------------------------------------------------------------------**

**-- Number of unique countries and athletes participating in each Summer Games**

SELECT

Year,

COUNT(DISTINCT NOC) AS NumberOfCountries,

COUNT(DISTINCT ID) AS NumberOfAthletes

FROM

athlete\_events

WHERE

Season = 'Summer'

GROUP BY

Year

ORDER BY

Year ASC;

-- Number of unique countries and athletes participating in each Winter Games

SELECT

Year,

COUNT(DISTINCT NOC) AS NumberOfCountries,

COUNT(DISTINCT ID) AS NumberOfAthletes

FROM

athlete\_events

WHERE

Season = 'Winter'

GROUP BY

Year

ORDER BY

Year ASC;

**Two new tables are Created Named 1. Country\_Stats table and populates it with data sourced from the World Bank. This provides a more accurate basis for correlating Olympic performance with national statistics.**

**2. Host\_Cities table and populates it with a comprehensive list of all modern Olympic Games hosts, Saved in a different SQL files.**

**-- Correlate medal count with population and GDP for a given year**

SELECT

ae.NOC,

COUNT(ae.Medal) AS TotalMedals,

cs.Population,

cs.GDP

FROM athlete\_events AS ae

JOIN Country\_Stats AS cs ON ae.NOC = cs.NOC AND ae.Year = cs.Year

WHERE ae.Year = 2016 AND ae.Medal IS NOT NULL

GROUP BY ae.NOC, cs.Population, cs.GDP

ORDER BY TotalMedals DESC;

**-- Analyze "host country advantage" by comparing medals won as host vs. non-host**

SELECT

ae.NOC,

(SELECT COUNT(\*) FROM athlete\_events WHERE Medal IS NOT NULL AND NOC = ae.NOC AND Games IN (SELECT Games FROM Host\_Cities WHERE Host\_NOC = ae.NOC)) AS MedalsAsHost,

(SELECT COUNT(\*) FROM athlete\_events WHERE Medal IS NOT NULL AND NOC = ae.NOC AND Games NOT IN (SELECT Games FROM Host\_Cities WHERE Host\_NOC = ae.NOC)) AS MedalsAsGuest

FROM athlete\_events AS ae

GROUP BY ae.NOC

ORDER BY MedalsAsHost DESC;

**-- Showing the evolution of average age for Gymnastics medalists**

SELECT

Year,

AVG(Age) AS AverageAge,

AVG(BMI) AS AvergaeBMI

FROM athlete\_events

WHERE Sport = 'Gymnastics' AND Medal IS NOT NULL

GROUP BY Year

ORDER BY Year ASC;

**--Identify which NOCs are dominant in which specific sports.**

**-- Find the top 5 countries in Fencing by medal count**

SELECT TOP 5

NOC,

COUNT(Medal) AS MedalCount

FROM athlete\_events

WHERE Sport = 'Fencing' AND Medal IS NOT NULL

GROUP BY NOC

ORDER BY MedalCount DESC;

**Time-Series and Era Analysis:**

**Divide the dataset into distinct historical periods to analyze shifts in medal distribution.**

**-- Compare medal distribution between two eras for Athletics**

SELECT

CASE

WHEN Year < 1980 THEN 'Pre-1980'

ELSE 'Post-1980'

END AS Era,

NOC,

COUNT(Medal) AS MedalCount

FROM athlete\_events

WHERE Sport = 'Athletics' AND Medal IS NOT NULL

GROUP BY

CASE

WHEN Year < 1980 THEN 'Pre-1980'

ELSE 'Post-1980'

END,

NOC

ORDER BY Era, MedalCount DESC;

--Hypothesis-Driven Querying:Answering the additional questions identified. (Female vs. Male):

-- Compare medal counts for men and women in Swimming for the USA

SELECT

Sex,

COUNT(Medal) AS MedalCount

FROM athlete\_events

WHERE NOC = 'USA' AND Sport = 'Swimming' AND Medal IS NOT NULL

GROUP BY Sex;

**-- Find the most successful decade for Italy in Fencing**

SELECT

(Year / 10) \* 10 AS Decade,

COUNT(Medal) AS MedalCount

FROM athlete\_events

WHERE NOC = 'ITA' AND Sport = 'Fencing' AND Medal IS NOT NULL

GROUP BY (Year / 10) \* 10

ORDER BY MedalCount DESC;

**Advanced Analysis: Next Steps**

This section outlines the next phase of the project, focusing on deeper statistical analysis, broader feature exploration, and the creation of new, insightful metrics.

Diving Deeper: Correlation and Regression

We will now move beyond descriptive statistics to investigate the relationships between variables.

**Creating New Metric 1.BMI to find relationships with medals won**

ALTER TABLE dbo.athlete\_events

ADD BMI DECIMAL(5,2);

UPDATE dbo.athlete\_events

SET

BMI = Weight/POWER(Height/100.0,2)

WHERE

Height IS NOT NULL AND Height > 0 AND Weight IS NOT NULL;

**--BMI vs. Medals Won**

SELECT

ISNULL(Medal, 'No Medal') AS MedalCategory,

AVG(BMI) AS AverageBMI,

COUNT(\*) AS NumberOfAthletes

FROM

dbo.athlete\_events

GROUP BY

ISNULL(Medal, 'No Medal')

ORDER BY

MedalCategory;

**--BMI vs. Medals Won by Season**

SELECT

Season,

ISNULL(Medal, 'No Medal') AS MedalCategory,

AVG(BMI) AS AverageBMI,

COUNT(\*) AS NumberOfAthletes

FROM

dbo.athlete\_events

GROUP BY

Season,

ISNULL(Medal, 'No Medal')

ORDER BY

Season,

MedalCategory;

**Calculating Medal Conversion Rate for the top 15 medal-winning countries (2nd New Metric)**

**Why: Simply counting total medals favors larger countries that send more athletes. The Medal Conversion Rate measures efficiency: (Total Medals Won / Total Participations). A high rate indicates that a country is highly effective at turning its Olympic appearances into podium finishes.**

SELECT TOP 15

NOC,

COUNT(CASE WHEN Medal IS NOT NULL THEN 1 END) AS TotalMedals,

COUNT(\*) AS TotalParticipations,

(CAST(COUNT(CASE WHEN Medal IS NOT NULL THEN 1 END) AS FLOAT) / COUNT(\*)) \* 100 AS MedalConversionRate

FROM

athlete\_events

GROUP BY

NOC

HAVING

COUNT(\*) > 1000

ORDER BY

MedalConversionRate DESC;

**3rd New Metric: "Athlete Versatility Index"**

**Why: Some athletes are specialists, while others display incredible versatility by competing in multiple sports or disciplines. This metric, calculated as the COUNT (DISTINCT Sport) For each athlete, it allows us to identify and celebrate these unique individuals.**

**-- Find the most versatile athletes who have competed in the most distinct sports**

SELECT TOP 10

Name,

Team,

COUNT(DISTINCT Sport) AS VersatilityIndex

FROM

athlete\_events

GROUP BY

Name, Team

ORDER BY

VersatilityIndex DESC;

**Advanced Analysis- Relationships/Correlation**

**--Compare the average height of medalists vs. non-medalists in Basketball**

SELECT

CASE

WHEN Medal ='NA' THEN 'Non-Medalist'

ELSE 'Medalist'

END AS Status,

AVG(Height) AS AverageHeight,

COUNT(\*) AS NumberOfAthletes

FROM

athlete\_events

WHERE

Sport = 'Basketball'

GROUP BY

CASE

WHEN Medal ='NA' THEN 'Non-Medalist'

ELSE 'Medalist'

END;

**--Medalist 191.126693685658 1080**

**--Non-Medalist 187.852372060629 3456**

**A clear positive correlation exists between an athlete's height and their likelihood of winning a medal in Basketball. Taller players have a distinct advantage, and this relationship is one of the strongest and most intuitive in the dataset.**

**--Age and Gymnastics Medals**

SELECT

CASE

WHEN Medal ='NA' THEN 'Non-Medalist'

ELSE 'Medalist'

END AS Status,

AVG(Age) AS AverageAge,

COUNT(\*) AS NumberOfAthletes

FROM

athlete\_events

WHERE

Sport = 'Gymnastics'

GROUP BY

CASE

WHEN Medal ='NA' THEN 'Non-Medalist'

ELSE 'Medalist'

END;

**Analysis: Age and Gymnastics Medals**

After analyzing the data to prove or disprove the negative correlation between age and winning medals in Gymnastics. Interestingly, the analysis of this dataset shows the opposite of the proposed hypothesis. The data indicates that the average age of medal winners in Gymnastics is slightly higher than that of non-medalists.

**Future Prediction (Linear Regression):**

The strong linear relationship between certain physical attributes and success in specific sports suggests that linear regression could be a valuable predictive tool. For example, we could build a model to predict the probability of winning a medal based on Age, Height, and Weight for a given sport.

This analysis could be performed in a statistical environment like Python or R, using the data extracted from our SQL database.

**What Jumps Out Now?**

The Event column is a rich, underutilized feature. Analysis so far has focused on the Sport level (e.g., "Athletics"), but drilling down into the specific event (e.g., "Athletics Men's 100 meters") will provide much more granular and actionable insights. We can analyze which specific events yield the most medals for each country.