SportsStats Project Proposal

Step 1: Preparing for the Project Proposal

Which client/dataset did you select and why?

The client I chose to work with was SportsStats. Performing sports analysis to provide key health insights for elite personal trainers intrigues me. As someone with a personal history with several sports such as basketball, martial arts, fencing, bicycling, and more, being able to perform analysis to gain insights on sports is something that speaks to me personally. Additionally, I reside within a culture that is heavily influenced by health and well-being as an ideal to pursue, so I find that this analysis that will be able to provide key health insights will provide value not only to the client I will be working with in this project, but also to the general public as well. The datasets provided by SportsStats consists of historical data spanning 120 years from the Olympic Games with over 269K records and 230 National Olympic Committees referenced. This dataset is particularly suited for SportsStats because it enables exploration of long-term performance indicators, demographic patterns, and region-level trends essential for storytelling and training insights.

Describe the steps you took to import and clean the data.

To import the data, I read the csv files provided using Python code in a Jupyter notebook as data frames using Pandas, created a connection between those data frames and a SQLite database, and uploaded SQL extensions to be able to perform SQL queries in Jupyter notebook. From there, I performed a thorough data quality assessment, looking at qualities such as accuracy, completeness, validity, consistency, timeliness, uniqueness, and reliability. Once the issues were identified through the assessment, data cleaning procedures were implemented to remove duplicates, impute missing values, correct inaccuracies, standardize formats, and manage outliers. This was done through, range checks, category checks, joins between tables, cross-referencing with Olympedia and official IOC documentation, and regex-based identification of inconsistent team names.

Perform initial exploration of data and provide some screenshots or display some stats of the data you are looking at.

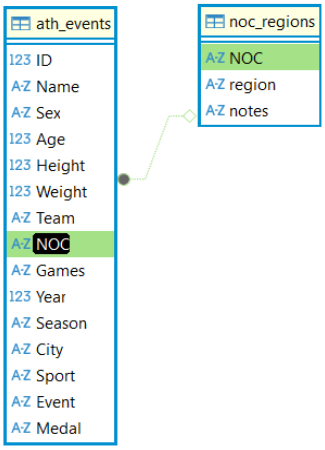
Early data exploration revealed:

- 269,661 records, spanning 1896–2016
- 135,571 unique athletes
- Significant missing biometric data before and during mid-20th century
- 51 Games total
- 66 sports and 765 distinct events
- Medal structure irregularities due to ties and team events
- Validity confirmed for most demographic and event-based fields

Screenshots/statistics (included in notebooks):

- Record counts
- Distinct counts per column
- Age distributions (From 10 to 97 years old)
- Height/weight ranges (127 to 226cm, 25 to 214kg)
- Temporal distribution of games (Discrepancies due to World Wars I and II)
- Geographical representations (How many teams represented different regions)
- Sports/event breakdown (For both Summer and Winter Games, also considering discontinued games and how many distinct events in each sport category)

Create an ERD to show the relationship of the data you are exploring:



There are two datasets used. One primary dataset consisting of athlete events (269K+ records) and one reference dataset consisting of NOCs and their associations with different regions. The relationship between both datasets lie within both NOC columns from each table, with a many-to-one relationship from ath_events to noc_regions.

Step 2: Develop Your Project Proposal

Project Description

SportsStats has requested an in-depth analysis of 120 years of Olympic history to uncover meaningful trends related to athlete demographics, national participation, and competitive outcomes. This project will leverage a cleaned version of the historical Olympic dataset to identify patterns across age, gender, sport categories, regions, and medal distributions. The resulting insights will be valuable for sports journalists developing data-backed narratives, athletic trainers seeking long-term performance benchmarks, and organizations evaluating historical and regional representation. By combining SQL-based data extraction with structured statistical exploration, the analysis will provide a comprehensive historical lens into how the Olympics (and its athletes) have evolved over time. The final deliverable will provide actionable findings and visualizations to guide SportsStats' media partners and performance analysts.

Guiding Questions

These questions intentionally reflect available data fields and the exploration that follows:

- How have athlete demographics (particularly age, gender, height, and weight) evolved over the 120-year history of the Olympics?
- Which regions and National Olympic Committees have shown the greatest growth, consistency, or decline in participation and medal performance over time?
- How do different sports compare in terms of competitiveness, athlete characteristics, and medal distributions across eras?

Initial Hypotheses

- Athlete demographics have become more standardized over time, with modern athletes displaying narrower age, height, and weight distributions due to professionalization and improved training science.
- Certain regions (e.g., Western Europe, USA) will show consistent dominance in medal counts, reflecting economic and infrastructure advantages, while others may emerge in specific eras.
- Sports with high physical demands (e.g., athletics, swimming) will have younger or more homogeneous athlete profiles, while historically non-physical events (e.g., art competitions) may show broader demographic ranges.

Approach

The analytical approach will begin by examining athlete demographic fields (Age, Sex, Height, Weight) to quantify their distributions across different decades and Olympic eras. Next, I will analyze region and NOC-level participation and medal outcomes using joins between the cleaned ath_events and noc_regions tables. I will explore potential relationships between age, sport type, event category, and medal performance to test whether demographic traits correlate with competitive outcomes. Temporal trend analysis will be used to evaluate how the Olympics have evolved over the past century, including shifts in gender participation and the introduction/discontinuation of sports. Throughout the analysis, SQL-based aggregation, window functions, and comparative statistics will be used to validate or refute the hypotheses. Final insights will be presented with supporting tables and visualizations derived from the cleaned dataset.