Olympics Data Analysis Project Proposal

Author: Bibhudendu Behera

Date: 18 October 2025

1. Project Description

This project analyzes the Olympics dataset to explore patterns, trends, and relationships among athletes, events, and medal outcomes over 120 years of Olympic history. The client, SportsStats, is a sports analytics firm seeking insights to highlight long-term Olympic trends and athlete demographics. The audience includes data scientists, sports journalists, and performance analysts interested in understanding gender participation, country performance, and evolving medal dynamics.

2. Questions

1. How have total Olympic medals evolved over time across different sports and countries?
2. How has gender participation changed throughout Olympic history?
3. Which sports and nations consistently dominate medal counts?

3. Hypotheses

1. Athletics and Swimming are likely to be the top medal-producing sports due to event diversity.
2. Gender participation has grown significantly over the past decades, especially post-1980.
3. Nations with higher GDP and consistent participation (like USA, USSR, and China) dominate medal counts.

4. Approach

To validate these hypotheses, the dataset will be analyzed using Python (Pandas, NumPy) for data manipulation, and Seaborn/Matplotlib for visualization. Grouping and aggregation methods will explore country and sport-level trends. Time-based visualizations will track medal evolution and gender ratios. Metrics include medal counts, gender distribution, and sport-level dominance. Findings will be presented through clear, interactive visualizations.

5. Data Preparation (Cleaning Steps)

The dataset contained missing values primarily in Age, Height, and Weight. Median imputation was applied to Age, while Height and Weight were retained as missing due to their limited impact on medal-based analysis. Data structure was verified using df.info() and df.isnull().sum(), ensuring dataset integrity without introducing synthetic bias.

6. Initial Exploration / Key Statistics

Key findings from the initial data exploration:

Total Records:	271,116 athlete entries
Time Span:	1896 - 2016 (120 years)
Gender Split:	~65% Male, ~35% Female
Total Countries (NOCs):	230
Number of Sports:	66
Top 3 Sports by Medal Count:	Athletics, Swimming, Rowing

7. Entity Relationship Diagram (ERD)

The ERD below illustrates how key entities (Athlete, Event, Team/NOC, and Medal) are interconnected in the dataset: