Gianna Burgess

003103859

Dr. Moussa Doumbia

Intro to Data Science

1 December 2025

120 Years of Olympic History

I. Introduction

This project analyzes 120 years of Olympic history (1896-2016) using athlete and NOC region datasets. The goal is to analyze trends, demographics, medal distribution, the impact of physical attributes on Olympic success, and overall participation.

II. Data Cleaning and Understanding

The raw dataset consisted of two files: athlete_events.csv, the primary data, and noc_regions.csv, the country mapping. The following steps were done in the cleaning process to ensure that the data was accurate and appropriate for analysis:

1. Merging Datasets
    a. The two datasets were merged by using a left join on the NOC code. In doing this, a singular data frame was created for better analysis and accessibility.

    b. After merging, the NOC column was dropped, as the team and region columns provided enough information. Missing entries in the region columns were imputed with the "team" names.

2. Duplicate Handling
    a. 1385 duplicate entries in athlete participation in the same events were identified and dropped for accurate counting of the unique participation and medals.

3. Missing Value Imputation
    a. The .isnull().sum() method was utilized to see how many null values were in each column. The results showed that the age, weight, height, and medal columns were the only ones with null values present, with the medal column having the greatest amount at 231,333 null values.
        i. Knowing this, the null values in the Medal column were imputed with "No Medal" using the .loc[] and .fillna() methods to find the missing entries in the column and replace them with the "No Medal" value.

ii. The missing values in the age, height, and weight columns were imputed using the median grouped by sex and sport; the median was chosen as it is a robust measure of central tendency. To do this, multiple methods were chosen, including .transform() to apply a function to each group—which was created using the .groupby() method—, lambda x: ... to represent the age column with "x" within each specific group, and x.fillna(median()) to calculate the median age of each group and use that value to fill any NaN value within said group's data.

4. Type Conversion
    a. The year and season columns were converted into categorical data for more efficient storage and processing.

5. New Column Creation
    a. The BMI column was created based off the BMI formula (Weight (kg) / [Height (m)]^2) and accounting for the height being in centimeters. This allowed for more specific analysis.

    b. The Age Group column was created by using bins to define boundaries for the age groups, labels with the actual age ranges for each respective bin, and the pd.cut() function to segment and sort data into bins. The right=True condition was used to show that the bins are right-inclusive, meaning that all values up to and including 15—not 0 though—would be included in the <15 label.

III. Exploratory Data Analysis
    a. The dataset includes 135,571 unique athletes, 205 countries, and 66 unique sports.
        i. This was revealed through utilizing the .nunique() method—after creating a dictionary called unique_counts—to count the number of unique observations in the id, sport, and region columns. For key, value in was used to iterate the items of the dictionary and print them as a f-string.

    b. Male participation (195,353) historically has dominated female participation (74,378), which was deduced by using .value_counts() to count the number of men and women who have participated in the Olympics.
    c. Athletics (38,624) and gymnastics (26,707) show the highest amount of participation, due to the large amount of events they offer.
        i. The .value_counts() method was used to find the top ten sports and .nunique() to find the unique total of events in the top ten sports.

There is overlap between the two lists, with swimming, gymnastics, cycling, wrestling, shooting, and athletics being in both.

d. The overall distribution of medals is nearly perfectly proportional, with Gold at 33.6%, Silver at 33%, and Bronze at 33.4%.
    i. Creating a list (medal_list) and checking the values in the medal column via the .isin() method allowed for only rows where said values were in the medal_list to be selected. The .value_counts() counted the frequency of every unique value in the selected medal column.

e. Olympic participation significantly increased from the first games in 1896 (380) to 2016 (13,688).
    i. To find the participation over time, .groupby("Year") was used to find unique values in the year column and then use .count() to count the number of non-null ID entries within each year group.
    Finally, .sort_index() was used to ensure that the data would be sorted in chronological order.

f. The median athlete is 25 years old, 178 cm tall, and weighs 73.5 kg.
    i. This was revealed by using groupby() to group medal groups by physical attributes (age, weight, height, BMI) and then calculating the median value for each physical attribute; the .sort_values(by="Age") was utilized to sort the results based on the values in the age column.

IV. Data Visualization and Interpretation

The analysis used 7 different data visualizations to explore trends:

1. Top 10 Countries by Total Medals
    a. A bar chart was used to identify top-performing nations; it confirmed that the United States has dominated the Olympics, followed by Russia and Germany.

2. Gender Participation Over Time
    a. A line plot was used to visualize gender trends, showing that male participation has historically dominated. However, women's participation in the Olympics has grown exponentially since the mid-1900s, with the gender gap ending in the past 20 years.

3. Age Distribution by Medal Type

a. A boxplot was used to compare medalist ages by medal type. It shows that the median age amongst all medal types is 25 years old, confirming that Olympic success in general is concentrated around a specific age group.

4. Pair Plot of Physical Attributes
    a. A pair plot was used to explore multivariate physical attributes, showing that there is a strong linear correlation between physical attributes (weight, height, and BMI) but not age and size; there is not a specific body type encouraged for each specific age group.

5. Medal Count by Sport
    a. A bar plot with hue was used to understand sport-wise performance and highlights the structural bias in total medal counts, as high-volume sports like athletics and swimming outnumber basketball despite being less globally popular.

6. Medal Distribution
    a. A pie chart was used to visualize the overall distribution of medal types and highlight the nearly perfect proportionality of medal distribution; this shows the accuracy and fairness of the recorded data.

7. Heatmap of Correlation Analysis amongst Year and Physical Attributes
    a. An interactive heatmap was used to understand the correlation between the year of the Olympic Games and athletes' physical attributes; there's an interdependence of physical attributes that is indicated by the positive correlations between weight and BMI, height, and weight. The average body type of Olympic athletes has remained consistent throughout 120 years of Olympic history—as seen by the negative correlation between year and the physical attributes—suggesting that the physical requirements have not shifted.

V. Insights and Generalizations

    a. Medal Dominance:
        i. The USA and Russia have dominated in terms of total medals won, as highlighted in the Top 10 Countries by Total Medals bar chart.

        ii. Sports (gymnastics, swimming, and athletics) that have more individual events have also dominated, as they allow for more medal-winning opportunities.

    b. Gender, Age, and Physical Attributes' Impact:

i. Medal success, regardless of type, peaks at the median age of 25.
ii. The gender gap is closing—as seen in the line plot highlighting Gender Participation Over Time—despite male participation being historically higher.
iii. Medalists are taller, heavier, and have a bigger BMI than non-medalists, regardless of gender.

c. Participation Trends:
    i. Total participation from the first games in 1896 to the last in 2016 has significantly increased, which can be attributed to the gradual globalization and inclusion of women in the games.

    ii. The correlation heatmap shows that the average athlete has not gotten taller, older, or heavier over 120 years of Olympic history.

d. Limitations:
    i. Missing data in some of the columns resulted in an imputation of values, which means some estimation was needed to draw conclusions.

    ii. Structural bias in total medal count towards high-volume events.

    iii. Total medal count favors countries with the longest history of Olympic participation and sports with the highest number of events.

VI. Additional Analysis and Findings

Three additional research questions were explored using supporting visualizations:

1. Who are the most decorated athletes in Olympic history?
    a. By grouping ID and name, as well as counting the total number of medal wins, it was revealed that Michael Phelps (swimming) is the most decorated Olympian with 28 medals. An accompanying bar graph was used to better visualize the difference in medal wins between the top ten most decorated athletes.

2. Which sports do young athletes dominate in?
    a. By grouping sport and finding the median age of medal winners, a bar graph was created to show the top five sports with the youngest median medalist age; young athletes dominate in croquet, diving, and short track speed skating.

3. Do host countries tend to win more medals?

a. To answer this question, (1) host country mapping, a new dictionary, had to be defined; (2) host rows were identified by creating a new column, Host_Identifier, through the zip () method to combine the year and city columns into tuple pairs to make the new column values match the keys in the host country mapping dictionary; (3) a new DataFrame was created and filtered; and (4) the overall average of medals per year and the host year medals were calculated, merged, and then compared.
    i. After that, an accompanying bar graph was created and confirmed the "host effect," where host countries usually had significantly more medal wins during their host year than in their overall average medal performance in other years.