Amelia Willmann, Katie Malan, Janet Lim

DS4300: Large-Scale Storage & Retrieval

February 27, 2026

Homework 3: Data Science with Document Stores

ANALYSIS:

Each member of our group was an avid fan and consumer of all things Olympics this February with the 2026 Milano Cortina Winter Olympics, so we decided to take a deep dive into their history from a data-focused point of view. We found a dataset via a project on Galaxy Training that was sourced from Olympedia. The original raw data contains 234,522 rows and 17 columns from the 1896 Summer Games in Athens through the 2022 Beijing Winter Olympics. The file was only available to download as a CSV, so we converted it to a JSON per the assignment instructions using AI. Before we converted it, we stripped the white space and converted the variables into the proper format to clean it. We also got rid of some irrelevant data about countries participating in the Olympics, as our analysis is mostly focused on gender differences. For our Mongo implementation, we decided we wanted 5 collections: athletes, countries, events, games, and results. We also fleshed out the specific qualities we wanted in each (e.g., the 'athlete' collection has athlete_id, name, sex, birth_year, birth_date, birth_place, height_cm, weight_kg, nocs (country they are representing, 3-character abbreviation), teams (country they are representing, full name), and events). We used our other collections to store information about the specific games and where they were held, and the results and medals of each games. It is important to note that we chose to focus on 'events' rather than 'sports' (i.e., 'Individual, Men's' vs 'Tennis') to give us a more detailed, gender-focused analysis.

The 2026 Milan Winter Olympics sparked discussions globally about the rise in women's sports. Since this dataset encompasses data from the first modern Olympics in 1896, we wanted to explore the patterns in female participation in the Olympics in the last 130 years. We expected there to be an increase. This first visualization (Figure 1) indicated that there was a stark difference in the number of female athletes between the Summer and Winter Olympics because of the zigzagging line. To explore this further, our next graph (Figure 2) breaks down the number of female athletes by season, displaying that the Summer Olympics saw a greater increase in the number of female athletes than the Winter Olympics. However, it is also important to consider that the Summer Olympics also has significantly more events for women than the Winter Olympics does. Nevertheless, the increase in the number of female athletes despite a somewhat consistent number of events offered for women is notable.

While discussing our project, our group became curious about what women's events were most popular overall and which were most popular in the dataset's most recent Olympics (2022 Beijing Winter). We were surprised by how popular cross-country skiing was in 2022, considering it isn't widely covered on TV outside of the Olympics. It was encouraging to see women participating in slalom and giant slalom, events that women were only allowed to

compete in in 1948 and 1952, respectively. For the Summer Olympics, it was exciting to see that the Marathon event was in the top 10 female athlete events (Figure 4), considering that the first woman to run an official marathon was Katherine Switzer in 1967 at the Boston Marathon. The high prevalence of gymnastics was expected considering it has six events (team, all-around, vault, uneven bars, balance beam, and floor exercise). The pattern of male athletes numerically-dominating the Olympics is echoed in Figure 5; eight of the top ten events with the most number of unique athletes are men's events, with one women's and one open event rounding out the set. While we gleaned significant insights, there is still a lot of analysis we could do to explore the gender gap in Olympic sports. A next step we could take to further our analysis would be to divide the growth by season (Summer/Winter).

Our group was also curious about how nationality shapes Olympic success, so we explored how different countries have risen and fallen as medal leaders over time. One of the most striking stories in the data was China's transformation from an Olympic outsider to a global superpower. China didn't win a single medal until 1984 after rejoining the Olympics following a decades-long withdrawal over the Taiwan dispute. From there, their medal count grew steadily, peaking at the 2008 Beijing Games, where they earned more medals than at any prior Olympics.

To understand what drove this rise, we broke down China's medals by sport over time (Figure 6). Rather than a broad increase across all disciplines, the data revealed that China's success was concentrated in a handful of sports. Diving, Weightlifting, Gymnastics, and Shooting spearheaded China's Olympic success. This pattern suggests a strategy of targeted investment in specific disciplines rather than widespread athletic development, which is a fundamentally different approach from historically dominant nations like the United States, whose medals are spread across dozens of sports. Diving in particular stood out as China has won more Olympic diving medals than any other country by a wide margin.

** Note: We misread the instructions and thought we needed 3-4 APIs, so each group member made one. We realized we only needed one after we had all already made ours, but we decided they all had a lot of value and that the analysis on them was worth sharing. As such, our analysis exceeds the one-page limit.

VISUALIZATIONS:

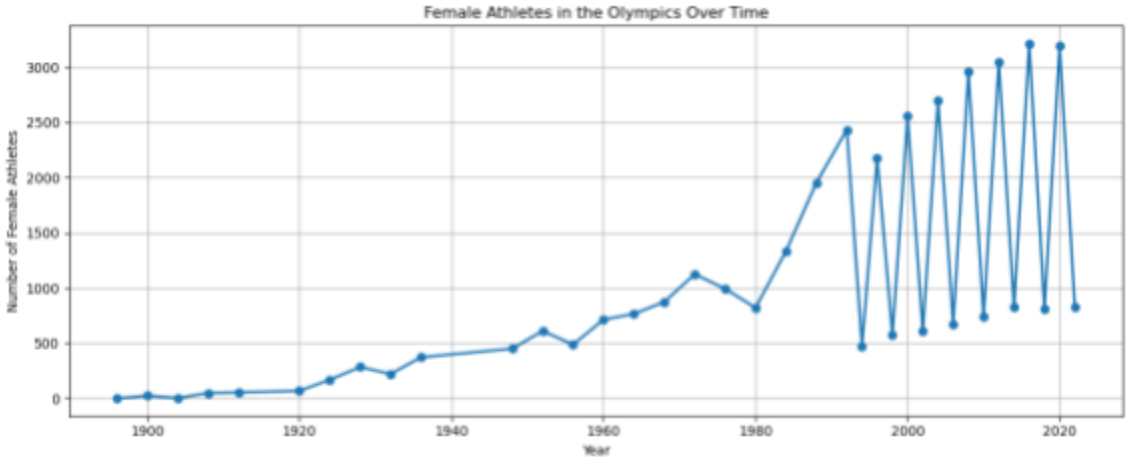

Figure 1: Female Athletes in the Olympics Over Time (AW)

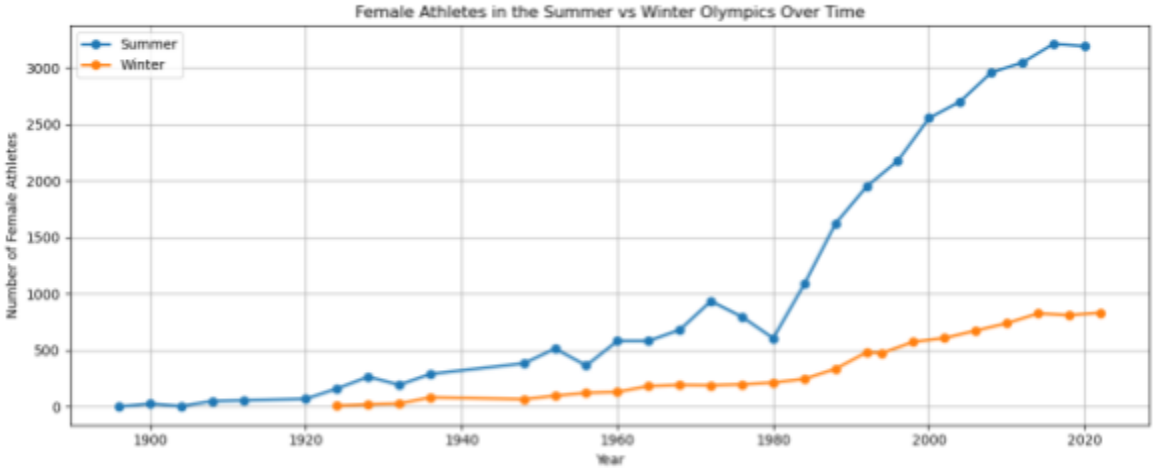

Figure 2: Female Athletes in the Summer vs Winter Olympics Over Time (AW)

Figure 3: Female Athlete Representation by Event (AW)

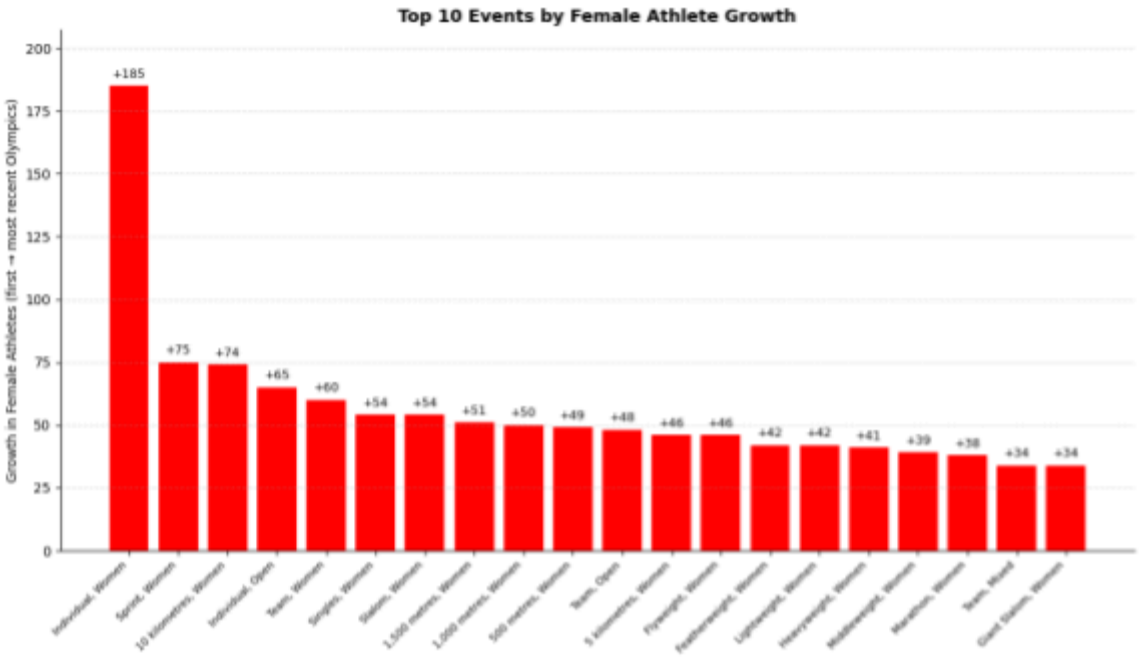

Figure 4: Top 10 Events by Female Athlete Growth (AW)

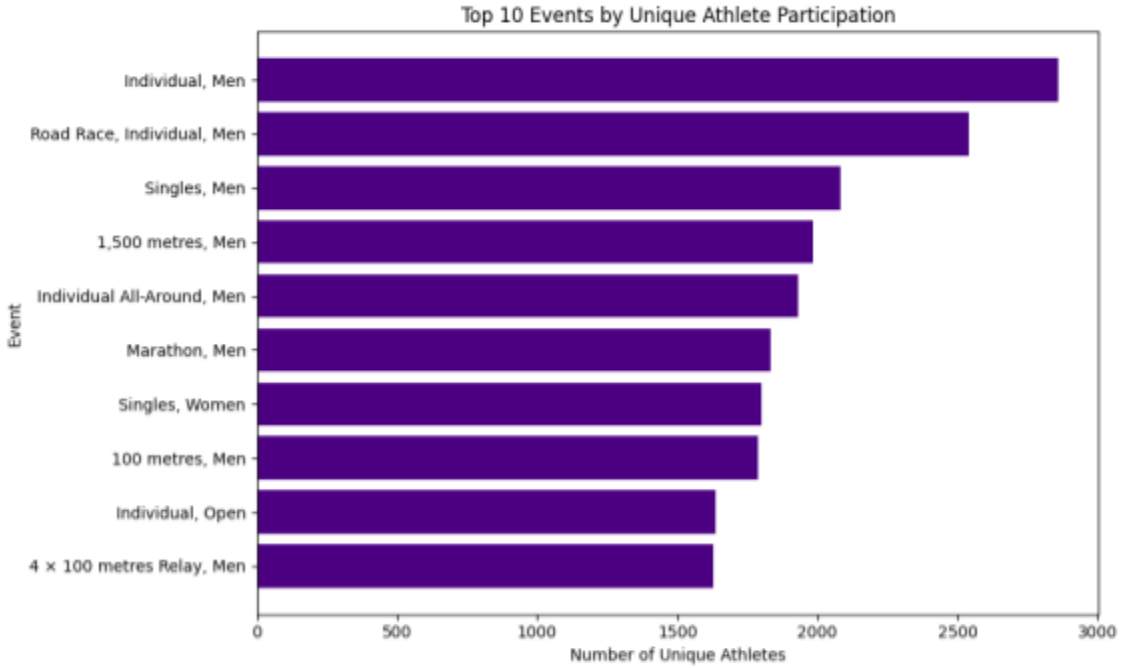

Figure 5: Top 10 Events by Unique Athlete Participation (KM)

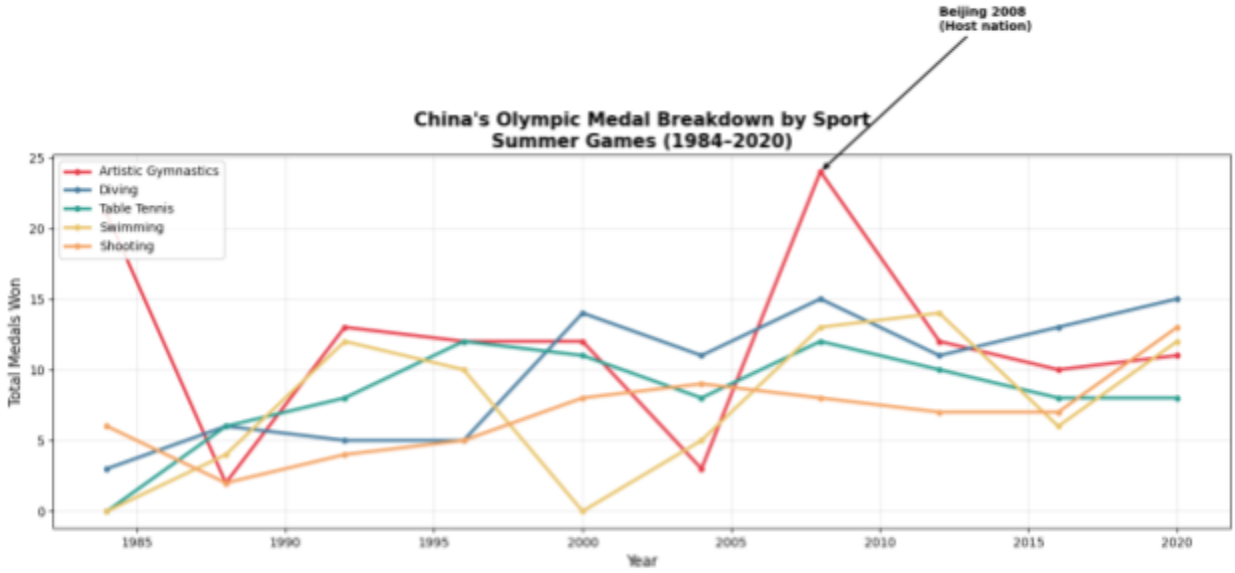

Figure 6: China's Olympic Medal Breakdown by Sport - Summer Games (1984-2020) (JL)