Ultramarathon Trends: Performance, Participation, and Gender Gap Analysis 2000-2022

B205306 and B231846

April 4, 2024

1 Overview

The recent rise in ultramarathon participation presents an intriguing field of study within data science, exploring the limits of human endurance and the influence of various factors on athletic performance. This report delves into the world of ultramarathon running, examining a dataset of race records from 2000 to 2022, with the aim of understanding the impact of event characteristics, demographic changes, and athlete lifestyles on the sport's evolution.

Analytical methods, including visualisations and statistical tests, have been employed to scrutinise changes in performance times across different race distances and participant demographics, particularly focusing on the narrowing gender performance gap. Additionally, patterns in global participation and the predictive modelling of performance times were analysed, considering variables such as event distance and athlete demographics.

The findings reveal a marked improvement in women's performance times and a general increase in international participation, underscoring the sport's growth and the potential of athletes. While the predictive model explains a substantial part of performance variability, it also underscores the multifaceted nature of endurance running, where factors such as training regimens and socio-economic contexts significantly impact outcomes. The insights from this study enrich the understanding of ultramarathon running, revealing a dynamic and evolving landscape shaped by a complex interplay of diverse influences.

2 Introduction

Context and motivation This data science study ventures into the realm of ultramarathon running, a sport that tests the limits of human endurance by pushing athletes to compete in distances surpassing the traditional marathon length of 42.195 kilometres including 100 mile races [1]. This study's focus on data spanning from 2000 to 2022 aims to uncover the recent evolutions within the sport, emphasising the shifting demographics and possible lifestyles of its participants. The surge in ultramarathons' popularity mirrors a broader societal trend towards challenging personal boundaries, both physically and mentally. This era is notably marked by significant strides in sports science, nutrition, and training techniques, revolutionising how athletes prepare for and excel in these gruelling competitions [2, 3].

By analysing the comprehensive dataset, the study intends to dissect the intricacies of athletic endurance and the dynamic participation landscape of ultramarathons. With a focus on recent decades, it aims to capture the contemporary essence of ultramarathon running, offering a lens through which to examine the interplay between athlete performance, and changing cultural narratives. Investigating these extreme endurance races offers valuable insights into human potential, the appeal of crossing conventional limits, and the impact of ultramarathons on personal and collective narratives of accomplishment.

Previous work In a recent 2024 BBC podcast exploring the performance of male and female athletes in ultramarathons, a study titled "The State of Ultra Running 2020" was discussed, highlighting intriguing findings. According to this study, the gap in average pace between men and women narrows as the

race distance increases, with men being 11% faster in marathons, but this difference dwindles to just 0.3% in 100-mile races. Intriguingly, for distances beyond 195 miles, women are purported to be around 0.5% faster than men. However, this claim is contrasted by the observation that men still hold faster world records in ultra distances, suggesting a complex interplay of factors affecting performance. This dichotomy presented on the podcast underscores the need for further research into gender dynamics in ultramarathon running, particularly considering the smaller number of women participating in these extreme distances [4, 5].

Objectives The study aims to explore the evolution of finishing times across various distances and demographics, revealing key performance trends in ultramarathon running. It will also identify shifts in global participation, highlighting emerging patterns. A predictive model for performance times, considering event distance and athlete attributes and demographics, is another core objective, alongside a focused analysis on gender dynamics to understand differences in participation and performance. To explore the dataset, we will be carrying out both statistical and visual analyses.

3 Data

Data provenance The dataset used for this study on ultramarathon running was authored by Elias Villiger and David Valero. It was made publicly available on Kaggle, where, in the Discussion forum, David mentioned that he received this dataset through a statistical analysis collaboration with the University of Zurich. The data was sourced from various public websites and is licensed under CC0 Public Domain [1]. This licensing implies (among other rights) that the dataset can be used for any purpose, including commercial uses, without the need for obtaining permission or giving credit [6]. The dataset was obtained directly from Kaggle in a CSV format, in compliance with the terms and conditions of the platform and the CC0 licence, ensuring that its use in this project is fully authorised and appropriate.

Data description The original dataset represents an extensive collection of ultramarathon race records spanning from 1798 to 2022, capturing over two centuries of endurance running history. It encompasses 7,461,226 records across 1,641,168 unique athletes.

Key columns in the dataset include the "*Year of event*", "*Event dates*", "*Event name*", "*Event distance/length*", "*Event number of finishers*", along with detailed athlete performance metrics such as "*Athlete performance*", "*Athlete club*", "*Athlete country*", "*Athlete year of birth*", "*Athlete gender*", "*Athlete age category*", "*Athlete average speed*", "*Athlete ID*". To ensure privacy compliance, original athlete names have been replaced with unique Athlete IDs.

The dataset categorises races by both distance (e.g., 50km, 100km, 50mi, 100mi) and length (e.g., 6h, 12h, 24h, 48h, 72h, 6d, 10d), covering a wide spectrum of ultramarathon formats - both distance based and time based. Additional data on athlete age, gender, and speed offer a comprehensive view for analysing performance trends, demographic shifts, and the evolution of ultramarathon running over time. This rich dataset serves as a foundational resource for in-depth exploration into ultramarathon participation and achievement, facilitating both historical analysis and predictive modelling endeavours.

Data processing In the process of preparing our dataset for analysis, we focused on ultramarathon records from 2000 to 2022, applying rigorous filtering to emphasise recent trends and developments within the sport. Our initial step involved categorising events into distance-based and time-based groups, acknowledging their unique aspects for targeted analysis. We then removed entries with "00:00:00 h" in 'Athlete performance', which signified either non-finishes or timing malfunctions, to ensure data integrity. For time-based events, for we standardised measurements of distances covered by athletes to kilometres, facilitating uniform analysis across varying event formats. Distance-based events underwent a conversion process, with athlete performance times transformed from diverse formats into minutes, creating a 'Performance Minutes' column for direct and comparative analysis. Additionally, we refined

gender data by removing any missing values, though these instances were minimal and had negligible impact on our study. We further organised our dataset by country for regional participation analysis, focusing on the top 10 countries after addressing a few missing 'country' entries.

These data processing steps included essential conversions and cleaning of missing values. Predictive modelling and gender analysis benefitted from these meticulous preparation steps, ensuring our study's findings are grounded in a clean, well-structured dataset ready for comprehensive analysis. We processed the data by applying multivariate linear regression, using gender, event distance and event year to predict performance in future events with an $R^2 \approx 0.65$.

4 Exploration and analysis

We used the rich source of data supplied regarding ultramarathon running to analyse finishing time performances for athletes for different types of events and shifts in regional participation patterns, predict better performance by looking at certain factors like gender and event distance and analyse the impact of athlete gender on performance times.

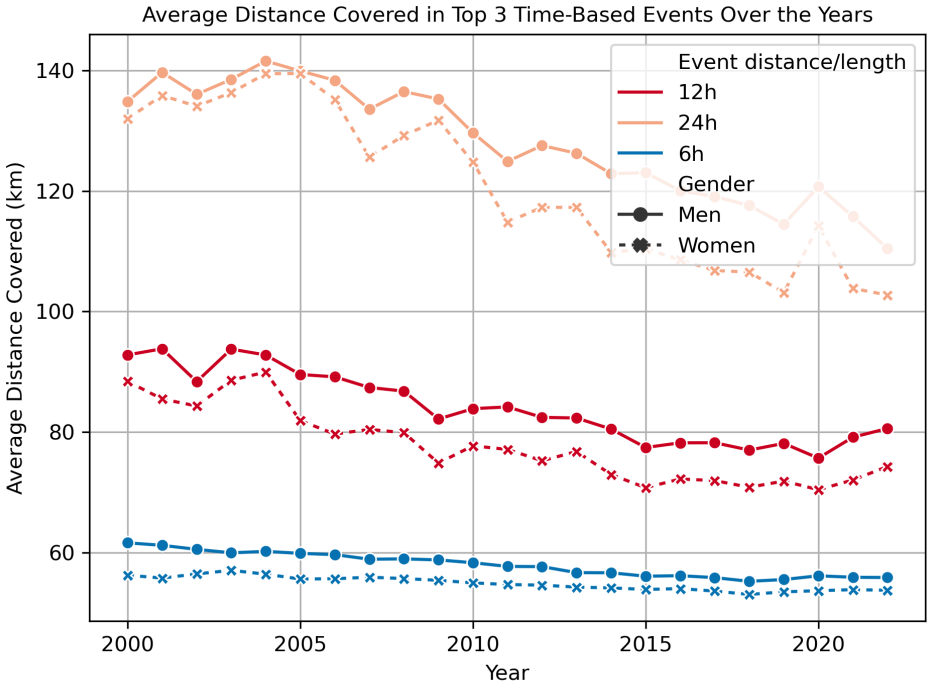

Figure 1: Line graph showing the average distance covered by men and women in the top 3 most participated in time-based events from 2000 - 2022

Ultramarathons all over the world are usually of two categories: either time-based (6h, 24h, 10h, etc.) or distance-based (50km, 50mi, 100km etc.). Analysing the data from ultramarathon events across two decades reveals evolving patterns in athlete performance and participation. Time-based events, as shown in Figure 1, illustrate men's consistent outperformance in distance covered compared to women, particularly in 24-hour events, which peaked in performance and then saw a decline, while shorter events like 6-hour races maintained more consistent results. This trend could be attributed to physiological factors that generally favour male endurance, as well as higher male participation rates leading to increased competition and performance levels. Despite a significant surge in participants for 6-hour events from countries like Russia indicated by Figure 2, this did not correlate with a proportional increase in average distances covered, suggesting varied skill levels among athletes. In contrast, Burma, while having fewer

participants, showcased top performances, possibly indicating a higher calibre of athletes or a more rigorous training and selection processes.

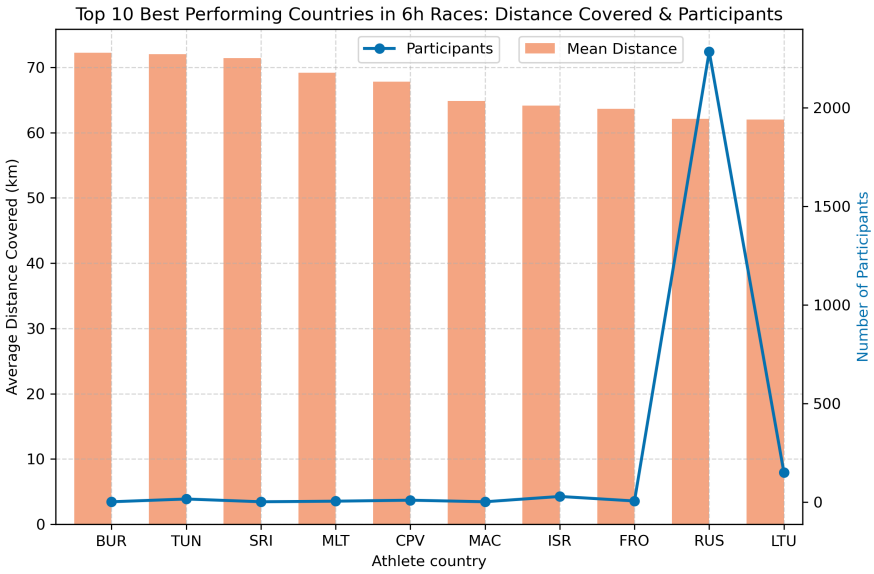

Figure 2: The top 10 best performing countries in 6h races along in terms of distance covered and a line plot showing the number of participants for each country.

For distance-based events, 50km, 100km and 50mi events constituted nearly 40.7% of the ultra-marathon distance-based events conducted from 2000-2022. In these, we observe an overall rise in average performance times across both genders, with particular variability noted among women in the 100km category around 2010. This could indicate fluctuating race conditions, training methodologies, or other external factors impacting athlete performance. Notably, the gender performance gap has been decreasing, especially in the 50km races, with a drop from 11.59% in 2000 to 6.28% in 2022, signalling a commendable improvement and that while men are still faster on average, women are improving at a rate that is closing the performance gap.

Furthermore, the 2019 data for 50km events (Figure 3) shows Lesotho (LES) with the highest number of participants yet not the best performance times, as opposed to Ethiopia (ETH), which had the fastest average times with fewer participants including the 2^{nd} fastest performer with 168.35 minutes at a 50km event in 2019. This discrepancy emphasises that a higher number of participants does not inherently result in superior performance; instead, possible factors such as athlete support systems, and environmental conditions play significant roles. Interestingly, countries leading in the 6-hour events (with more than 1750 participants) did not necessarily have the best performance times in the 50km events, implicating the importance of event distance which clearly dictates performance times.

To further analyse the influence of athlete country, we studied regional shifts in participation patterns. The top 10 countries in participation (Figure 4) likely owe their status to many factors including substantial investment in sports infrastructure and training, cultural emphasis on athletic participation and supportive government policies [7]. Additionally, these nations may benefit from larger populations, economic resources that enable widespread access to sports, and educational systems that promote athletic development, all contributing to a greater pool of competitive athletes.

There is a general upward trend in international participation, with the USA consistently leading, indicative of robust infrastructural support for the events in question. There is a dramatic surge of about 250% in China's participation post-2015 ,causing it to overtake 5 countries. This potentially reflects increased sports investments and the impact of upcoming international events such as the 2022 Winter

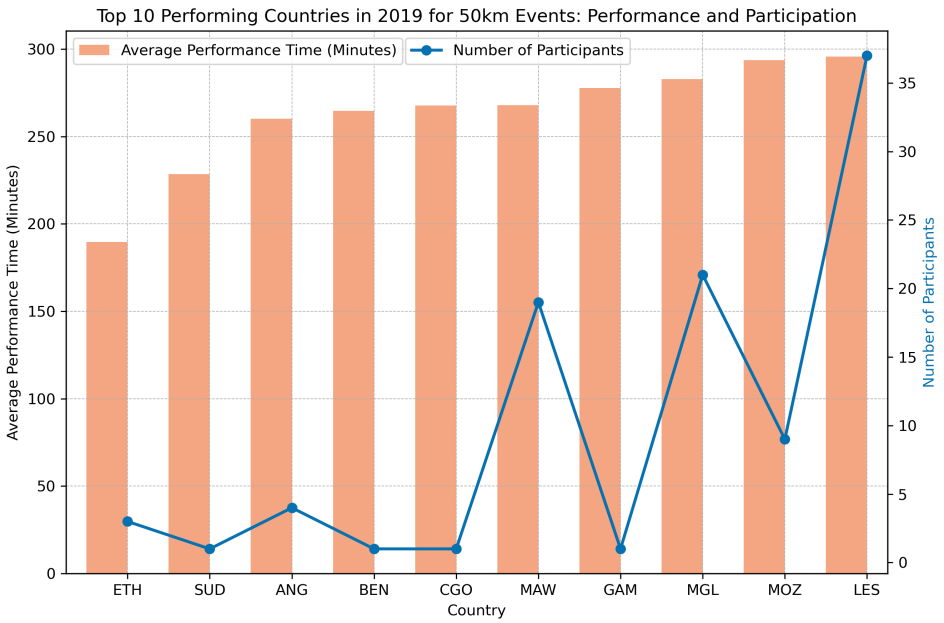

Figure 3: The top 10 best performing countries in 50km ultramarathon events in terms of average performance times and a line plot showing the number of participants for each country.

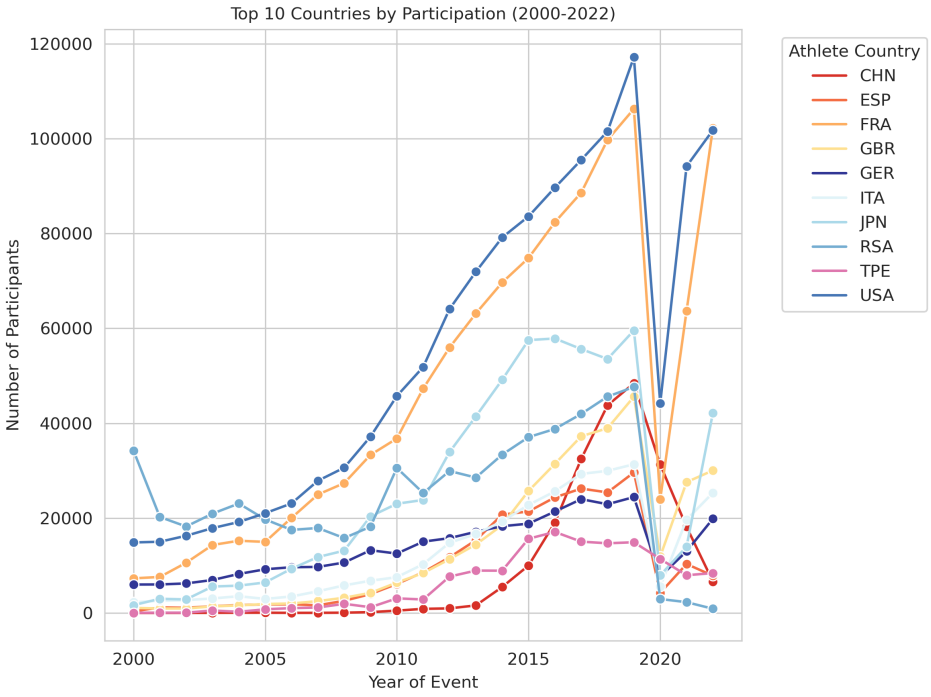

Figure 4: Top 10 countries having maximum number of participants from 2000 to 2022

Olympics [8]. The sharp decline across all countries around 2020 aligns with the advent of the COVID-19 pandemic, causing widespread event cancellations and travel restrictions. The pandemic's disruption of training schedules, and the overall health impact on athletes led to a noticeable decline in participation. However, Japan presents an outlier with a notable increase in participation post-pandemic (about 300%), potentially due to the rescheduled Tokyo 2020 Summer Olympics.

The gradual increase in participants from 2000 to 2010, and more notably the steep rise from 2010 to 2019 (as visible by the changing slope in Figure 4), suggests a growing engagement with the events, possibly driven by wider access to advanced training technologies, which have likely improved training efficiency and performance times.

Apart from event distance and athlete country, various event characteristics and athlete attributes affect performance times. An analysis into how gender affects performance will be explored below, broader social and biological insights, revealing how gender-specific factors, such as hormonal differences, sociocultural influences, etc. contribute to the observed performance outcomes. For this, we have focused mainly on the 50 km events, which saw maximum participation numbers.

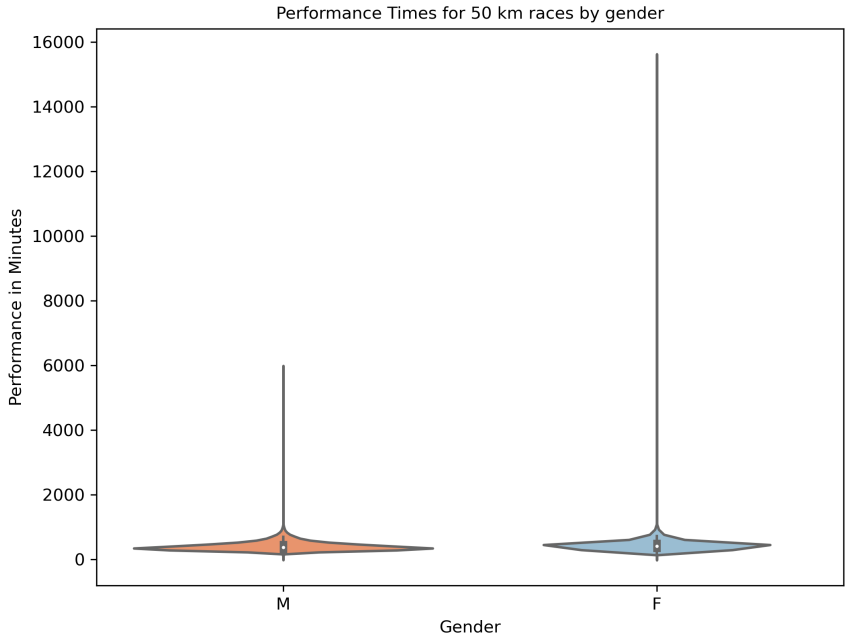

Figure 5: Performance Times for 50km races by Gender

Our Null Hypothesis H_0 posits that there is no significant effect of athlete gender on outcomes of performance times. Conversely, our Alternative Hypothesis H_1 suggests that gender does play a role in shaping performance times for these endurance events. The dataset for the 50 km races was divided into two groups - males (group 1) and females (group 2). The t-test was applied to determine if there is a significant difference between the mean performance times of the two groups and how they are related. The resultant t-statistic is -143.24, indicating that the mean of group 1 is lower than the mean of group 2 , indicating that on average males have lower performance times hence are faster than females for 50 km races. The p-value produced is 0.0 (rounded from a very small value close to zero) which is lower than 0.05, providing strong evidence against the null hypothesis, thus disproving it. Further, Figure 5 shows that the median for male performance times is lower than that for females, suggesting that males tend to complete 50 km races faster than females.

As seen in Figure 5, the distribution for female performance times is wider than for males, implying greater variability in the times of female runners. This suggests that female runners have a broader range of performance outcomes than male runners, possibly due to physiological factors or difference in participation rates. Both distributions have long tails extending to very high performance times, but

especially for females. These could be outliers or may reflect a subset of participants with significantly longer race completion times.

Since we recognised that performance times can be influenced by several complex factors, we introduced a multivariate linear regression model which considers both gender and distance among the variables. For the model we used a subset of the dataset which encompasses records from the top three distance-based events (50km, 50mi and 100km) spanning from the year 2000 to 2010. We focused on several independent variables: 'Year of event' which allows us to account for temporal trends in performance; 'Event distance/length' which provides insights into the relationship between race length and performance times; and 'Gender_M' which is a binary variable derived from one-hot encoding the 'Athlete gender' attribute, distinguishing male (1) from female (0) participants.

The dependent variable 'Performance Minutes' represents the time it took each athlete to complete the event, offering a direct measure of athletic performance. By training a Linear Regression model with these variables, we aimed to quantify the influence of time, event distance, and gender on performance.

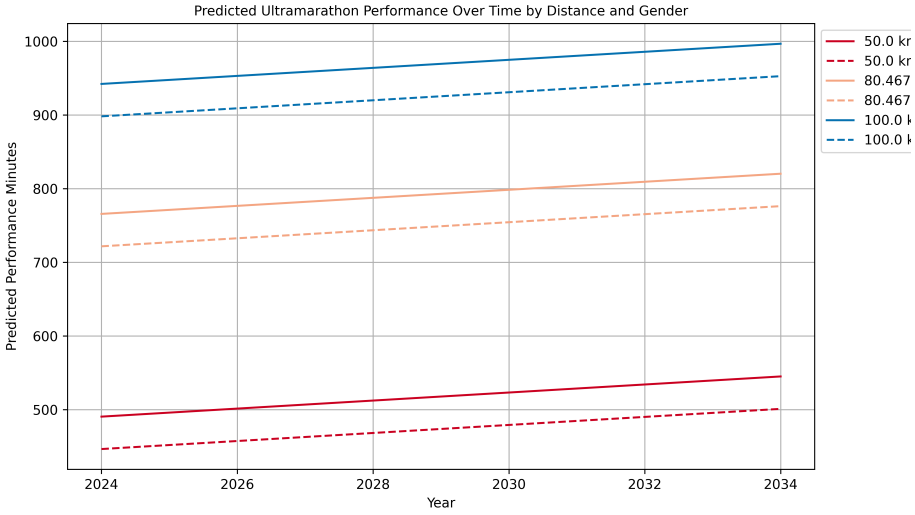

Figure 6: Predicted Ultra Marathon Finishing times from 2024 to 2034 varied by distance and gender

The linear model is represented as:

$$\text{Performance Minutes} = \beta_0 + \beta_1 \cdot \text{Year of event} + \beta_2 \cdot \text{Event distance/length} + \beta_3 \cdot \text{Gender\_M} + \varepsilon \quad (1)$$

In this equation:

- β_0 represents the intercept of the model.
- β_1 is the coefficient for the "Year of event" variable.
- β_2 is the coefficient for the "Event distance/length" variable.
- β_3 is the coefficient for the "Gender_M" variable, indicating the impact of being male on performance times.
- ε represents the error term of the model.

The produced $R^2 \approx 0.64690$ suggests that approximately 64.69% of the variance in performance times can be explained by the model's inputs. This is a moderately strong score, indicating that the model has a good ability to predict performance times based on the factors considered. We then used this model to predict future finishing times for varying by distance and gender as seen in Figure 6.

5 Discussion and conclusions

Summary of findings Time-based events showed that men generally outperformed women, with peak performances observed in longer events before a decline, while shorter events displayed more consistent results across genders. This underscores physiological and competitive differences, highlighting the endurance edge men usually have [9]. Distance-based events (50km, 100km, and 50mi races, accounting for nearly 40.7% of such events) showed an overall increase in average performance times for both genders. Notably, the gender performance gap in 50km races has been narrowing, indicating significant improvements among female athletes and a closing performance gap with male counterparts.

Regional participation patterns shifted greatly, with a significant rise in global engagement, particularly from the USA and China, the latter showing a 250% increase post-2015. The COVID-19 pandemic caused a temporary decline in participation, but Japan demonstrated resilience with a post-pandemic increase.

A deeper look into gender effects on performance through statistical analysis (t-test) provided strong evidence against our null hypothesis that gender does not affect performance times, particularly in 50km events. This was further supported by a multivariate linear regression model, which included gender, event distance, and year of event as variables. The model, explaining about 64.69% of the variance in performance times, underscores the significant impact of these factors on athlete performance. This comprehensive analysis not only highlights the evolving landscape of ultramarathon participation and performance but also underscores the narrowing gender performance gap, signalling a promising trend towards parity in endurance sports.

Evaluation of own work: strengths and limitations A strong aspect of our work is that it has broadly classified the large dataset of ultramarathons into time-based and distance-based events. We have also done both visual and statistical analyses to draw our conclusions. Using a multivariate linear regression model to predict future finishing times was the right approach as ultramarathon performance is affected by a variety of factors. However, the Mean Squared Error is 23037.11 and Root Mean Squared Error is 151.77 indicating that our model does not capture all the factors that cause year-to-year variability such as athlete age, training regimens, diet, weather conditions, course topography, and other socio-economic factors. We were not able to factor in the athlete age provided in the dataset for any analysis as 7.8% of the values were missing, which would lead to a significant data loss.

Comparison with any other related work As mentioned earlier, a 2024 BBC podcast delved into gender performance in ultramarathons, discussing a study which found that while men are generally faster by 11% in marathons, this gap almost vanishes to 0.3% in 100-mile races, and intriguingly, women outpace men by 0.5% beyond 195 miles [5]. Although we found no evidence to prove that women would outpace men, we did find data that supported improvement of women's pace in events in recent years.

Knechtle and Nikolaidis (2018) [2] found that men prove to be faster than women in ultramarathon racing, but the sex difference in performance has decreased in the past few years to 10–20% depending upon the length of the ultramarathon. This concurs with our findings regarding improvement of performance of women, especially in 50 km races. Further, Coates et al. (2020) [10] found that the mean finishing time for 50km ultramarathons was 6.4±1.5h for women and 6.2±2.3h for men which concurs with the mean finishing values that we found as well.

Improvements and extensions To improve the predictive power of the model, more variables affecting ultramarathon performance should be considered such as weather conditions, course topography, athlete training regimens, etc. Further, to extend the analysis, external datasets can be used to provide the missing age values. Moreover, conducting qualitative studies or surveys among ultramarathon runners could provide deeper insights into factors affecting performance and participation, such as motivation, perceived barriers, and personal strategies for training and competition. As an extension, we could use biometric data such as heart rate, elevation change, etc. from wearable technologies to find a correlation between these elements and performance times, making our findings more accurate.

References

[1] E. Villiger and D. Valero. *The big dataset of ultra-marathon running*. Accessed: 30 Mar. 2024. 2023. URL: https://www.kaggle.com/datasets/aiaiaidavid/the-big-dataset-of-ultra-marathon-running/data?select=TWO_CENTURIES_OF_UM_RACES.csv.

[2] B. Knechtle and P.T. Nikolaidis. "Physiology and pathophysiology in ultra-marathon running". In: *Frontiers in Physiology* 9.364 (2018). Retrieved on 2nd April 2024. DOI: https://doi.org/10.3389/fphys.2018.00634.

[3] M. Kazimierczak et al. "The Impact of Modern Ultramarathons on Shaping the Social Identity of Runners. The Case Study of Karkonosze Winter Ultramarathon". In: *International Journal of Environmental Research and Public Health* 17.1 (2019). Retrieved on 2nd April 2024. DOI: https://doi.org/10.3390/ijerph17010116.

[4] C. Stout and P. Ronto. *The State of Ultra Running 2020: Innovation in 'Data-Athletics' with Paul Ronto*. Retrieved 1 Apr. 2024]. 2020. URL: https://www.academia.edu/42483184/The_State_of_Ultra_Running_2020_Innovation_in_Data_Athletics_with_Paul_Ronto.

[5] Hartford T. (Presenter) PresenterName et al. *More or Less: Behind the Stats - Ultramarathons: Are women faster than men?* Retrieved on 30th March 2024. 2024. URL: https://www.bbc.co.uk/sounds/play/p0hg2764.

[6] Creative Commons (2019). *Creative Commons — CC0 1.0 Universal*. Retrieved on 1st April 2024. 2019. URL: https://creativecommons.org/publicdomain/zero/1.0/.

[7] Y. Noguchi, C. Kuribayashi, and T. Kinugasa. "Current state and the support system of athlete wellbeing in Japan: The perspectives of the university student-athletes". In: *Frontiers in Physiology* 13 (2022). Retrieved on 1st April 2024. DOI: https://doi.org/10.3389/fpsyg.2022.821893.

[8] B.E. Ainsworth and J.F. Sallis. "The Beijing 2022 Winter Olympics: An opportunity to promote physical activity and winter sports in Chinese youth". In: *Journal of Sport and Health Science* (2021). Retrieved on 2nd April 2024. DOI: https://doi.org/10.1016/j.jshs.2021.09.005.

[9] C. Hubble and J. Zhao. "Gender differences in marathon pacing and performance prediction". In: *Journal of Sports Analytics* 2.1 (2016). Retrieved on 2nd April 2024, pp. 19–36. DOI: doi:https://doi.org/10.3233/jsa-150008.

[10] A.M. Coates et al. "Physiological Determinants of Ultramarathon Trail Running Performance". In: *SportRiv* (2020). Retrieved on 1st April 2024. DOI: https://doi.org/10.31236/osf.io/y2kdx.