Analysing Trends in Ultramarathon Running

Francesco La Rosa, Víctor Baeza Hirnyak and Joseph Luke

April 2, 2024

1 Overview

This report seeks to uncover and analyse trends behind ultra-marathon running. We started by exploring changes in ultra-marathon participation over the past half-century. Our findings indicate that over the last fifty years, participation in ultra-marathons has increased exponentially, with the exception of the COVID-19 years (2020-2021). We then examined how performance, measured as average speed, is influenced by athletes' characteristics such as gender and age. We found that both of these characteristics have an impact on performance: with males exhibiting a higher average speed across all the analysed race categories and the 31-40 year age group being the fastest across most of these categories. We then developed a prediction model using Ordinary Least Squares (OLS) linear regression to predict performance, which initially had an R^2 of 0.149. Later, we improved the model by restricting our analysis to the top five events for our chosen race lengths, achieving an improved R^2 of 0.498. Finally, we analyzed the impact of experience—measured as the number of times an athlete has participated in a particular type of race—on performance. We found that experience corresponds to higher average performance, except in longer races.

2 Introduction

Context and motivation People have always pushed themselves to extreme limits and sought to test the boundaries of human endurance. One such way they have tested these limits is through ultramarathon running. Unlike traditional marathons, ultramarathons are a type of foot race that are a more intense form of traditional marathons which exceed a distance of 42.195km (or 26 mi), pushing the involved athletes to demonstrate exceptional endurance capabilities. There are two types of ultramarathons - those which have a predetermined distance, and those with a specified time, the goal of the latter being to cover as much distance as possible in said time [1].

Running as a sport dates back millenia, with some being documented from as far back as Ancient Egypt [2]. Some of the oldest foot races were held in Scotland, too - the Red Hose Race at Carnwath, which is a tradition that has lasted for over 500 years now since 1508 [3]. The dataset also spanning such a vast timeline encapsulates the fact that races in general have had a rich history. As such, exploring this data would allow us to examine how participants' performance may have changed over the years. This study not only aligns with the deep historical roots of running, but also addresses questions about how ultramarathon races have evolved worldwide.

Previous work As old as running competitively is, the sport of ultramarathoning has gained a significant increase in the number of participants in recent years, with a 345% observed increase from 2008-2018 [4]. Not only that, but a study was conducted that concluded that the number of younger (< 19 years old) participants have been increasing as well, with European and North American runners being the most responsible for this rise. The same study also concluded that the boys were overall faster than the girls, except for those born in Europe and Oceania [5]. Finally, a survey determined that the average age of participants seemed to be in their 40s, with a majority of them being married men who had a higher-level education [6].

Objectives As mentioned, ultramarathons have experienced a surge in popularity in recent years. The extent of this popularity measured over a longer period of time and how this sport has evolved should be studied. This report will investigate how both participation and performance in ultramarathons around the world has changed during the last few centuries, as well as whether or not there have been any significant shifts in where the participating athletes come from – whether there is a change in regional participation patterns.

Additionally, the report will seek to identify whether certain athlete's attributes, as well as the race category and season the race takes place in will have a significant effect on their performance and whether it provides them with an advantage, and then making a prediction model for any possible future athletes wishing to participate in similar events based on the already available data. We will then try to account for certain variables that may have not been initially present in the original dataset that might have affected the average speed, such as geographical data. Finally, we will analyse how another factor: the experience of the athletes (defined by the number of times they participated in an event for a particular race category) affects the average speed.

3 Data

Data provenance. The data [7] used for our report comes from David Valero and Elias Villiger, who aggregated ultramarathon statistics from websites in the public domain and posted them on Kaggle under the "aiaiaidavid" [8] username. The dataset was downloaded via a CSV file, and is available under a "CC0: Public Domain" licence, meaning that the provided dataset is able to be copied and modified "without asking for permission" [9]. A link to both the user's profile as well as the dataset is available.

Data description. The provided dataset contains 7,461,226 ultramarathon data entries documented between 1798 to 2022. There were 10 columns relevant for the analysis: year of event, event dates (where seasonal data was extracted), event name (information of where the race took place), event distance/length (describes the race category), event number of finishers, athlete country, athlete gender, athlete age category (what age category each athlete belonged to), athlete average speed (provided in km/h), and athlete ID (which anonymises but keeps track of each individual athlete's participation of races).

Data processing. A couple of measures have been taken to make the analysis of the data both easier and faster. To begin with, the columns have been renamed to shorter yet descriptive names. Furthermore, irrelevant information has been dropped. This includes removing the gender information present in the age column, as there already exists a gender column; and only considering the 11 most popular race lengths as suggested by the dataset author. Although such a decision erases around half of the data, it has been done so that we can focus on the most relevant races, without being cluttered with too many categorical values. On top of that, it makes running the rest of the analyses faster and avoids lengthy string parsing.

Where appropriate, columns have been casted from strings to their most efficient data types; such as integer for age or float for average speed. Finally, impossible and null values are temporarily dropped for each analysed question. This is done to avoid the loss of information caused by erasing an entire row, which contains just a single wrong value. For example, when analysing the average speed by age group, speeds above 21 km/h are deemed to be impossible and are thus dropped, as they are faster than the marathon world record [10]. In addition, rows with null values for either average speed or age are also dropped.

4 Exploration and analysis

Following the introduction's explanation about the recent growing interest in ultra marathons, we aim to further understand this trend. To that end, the report will examine the regional patterns of the participants' origins, as well as any trend on the number of participants across the years. The approach has been to count the number of participants from each continent for every year, and later find the relative proportion that those participants represent. Given that 99.7% of the data has been recorded in the past 50 years, the analysis starts in 1972, avoiding noise due to the lack of consistent data before that year. The relevant information is shown below, in Figure 1.

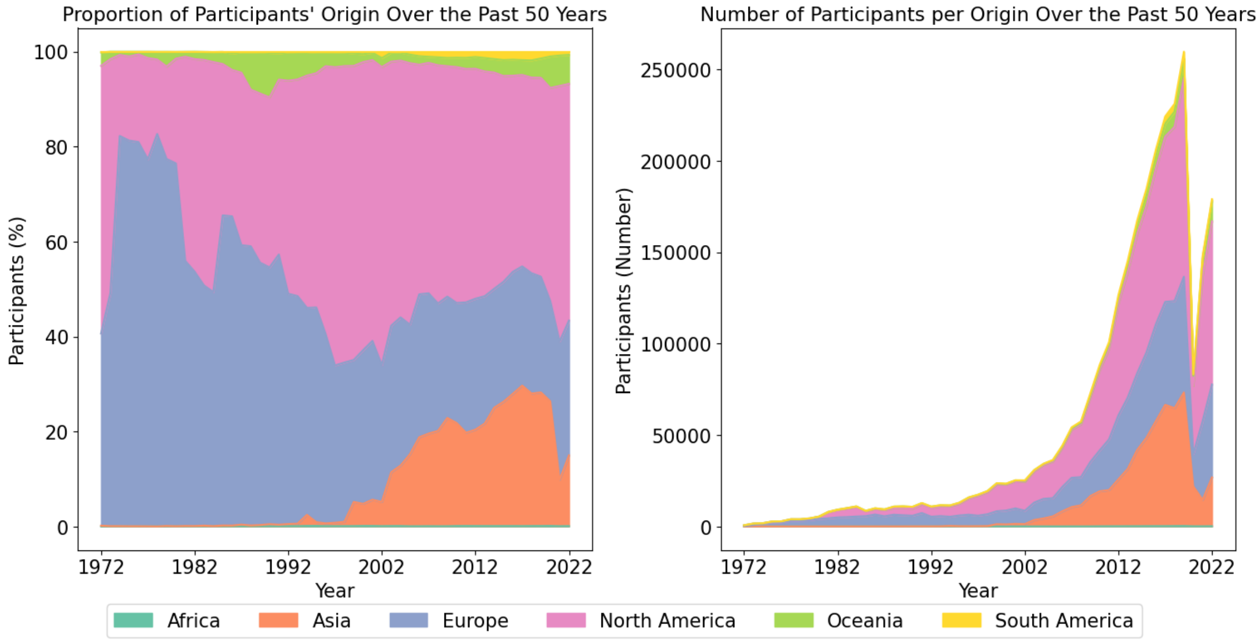

Figure 1: Variation of the participants' continent of origin over the past 50 years. The left graph represents the relative percentage (where the areas are stacked on top of each other) and the right graph shows the absolute number of participants' origin.

Some trends are worth noting for the left graph, such as the ones for Europe, North America and Asia. Europe started being the most popular continent of origin peaking at around 80% in the 1980's, and then slowly declined until reaching around 25% as of 2022. The proportion of runners from North America has been mostly increasing over time, reaching 50% in more recent years. Finally, Asian runners went from not participating at the start to a peak of 30% in 2019, finishing with a 20% representation of the total runners. The rest of the continents stayed at relatively low proportions throughout the entire time frame. The figure on the right also provides unique insights, showing a 55 fold rise in the absolute number of participants across all continents; which went from a mere 3,500 to 197,000 participants a year over the span of 50 years.

Looking at the graphs in Figure 1, some interesting features arise: the most notable being the COVID-19 pandemic, which accounts for a 65% drop in participation due to generalised lockdowns and movement restrictions. The left graph shows how the most affected runners were Asian, probably due to the zero-covid tolerance policy of some governments such as the Chinese or the Vietnamese ones. On the other hand, it is also noteworthy to comment about the peak in European participants around 1975; which does not have a clear cause, but might be explained by a decrease in North American participation due to the 1973-1975 recession. Finally, the figure hints at some features that could be explained by the globalisation of this sport: a higher diversity in the runners' origins (instead of being concentrated in Europe and North

America), and a popularisation of the sport, increasing the number of participants overall.

In the subsequent investigation, we will analyse athlete performance by focusing on the average speed of each athlete. The justification for this approach is to standardise for the two different types of races (fixed distance or fixed time) and the different possible lengths/times for each type of race present in the dataset (for example, 50km, 50mi, etc.). Longer races will of course take a longer time to complete, and vice versa, so the ratio between the distance covered and time taken will approximately be constant.

As can be seen in Figure 2, the average speed for male athletes is consistently higher than those for females, aligning with the findings from previous marathon analyses [11]. This trend can be observed for both the median and quartile values across all race categories. As expected, shorter lengths (50 km and 6h) have a higher average speed as the athletes do not need to pace themselves as much. On the other hand, progressively longer races have lower average speeds, especially for extreme races such as 6 and 10-day races where the participants need to pace themselves more and have a greater overall endurance in order to be successful.

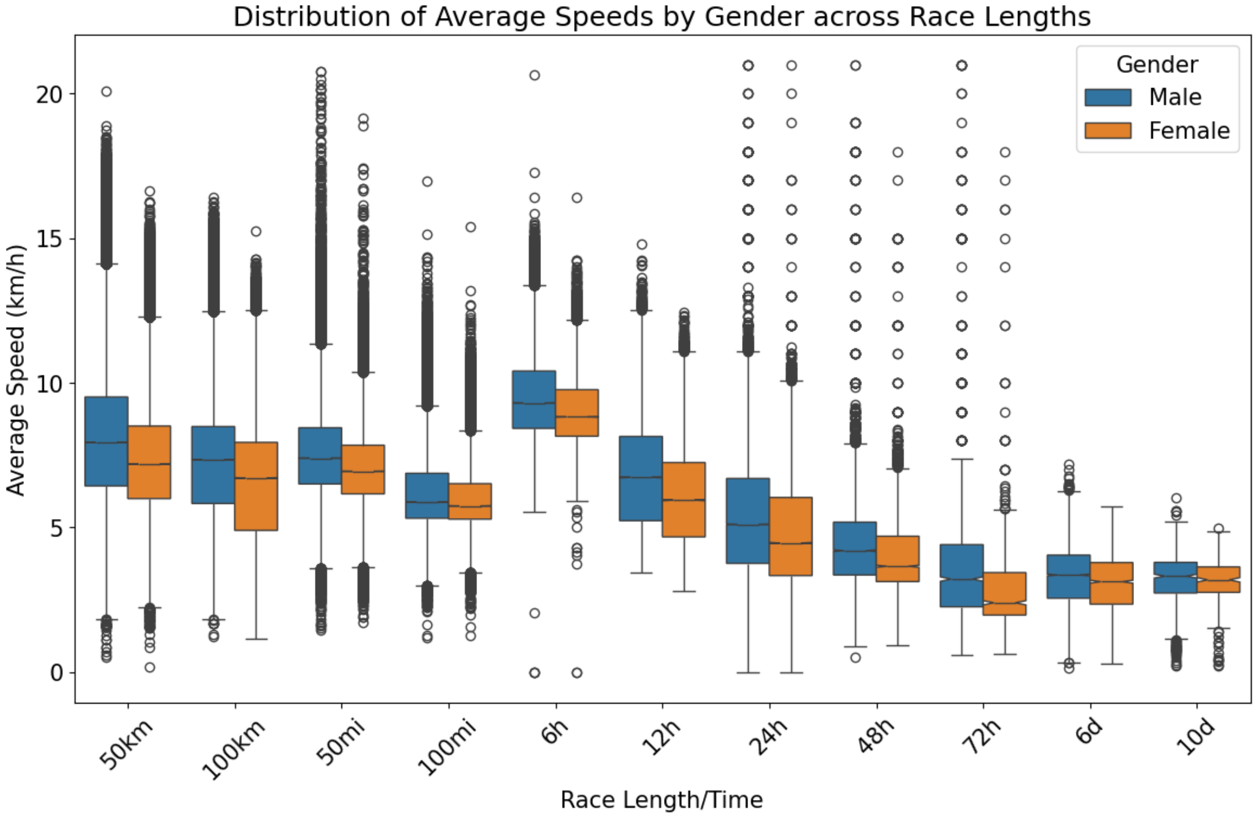

Figure 2: Average speed by gender across the most popular type of race categories.

A similar analysis can be done to the age groups as shown in Figure 3. For most categories, the 31-40 age group exhibits the highest median and quartile average speeds, closely followed by the 19-30 and 41-50 age groups respectively. Notably, the 19-30 age group is generally the second-best performing one for fixed length categories (with a notable exception for 100 km races), while the second best performing group for timed categories were 41-50. Finally, the worst performing age group is the one for ages above 70 as expected due to their old age.

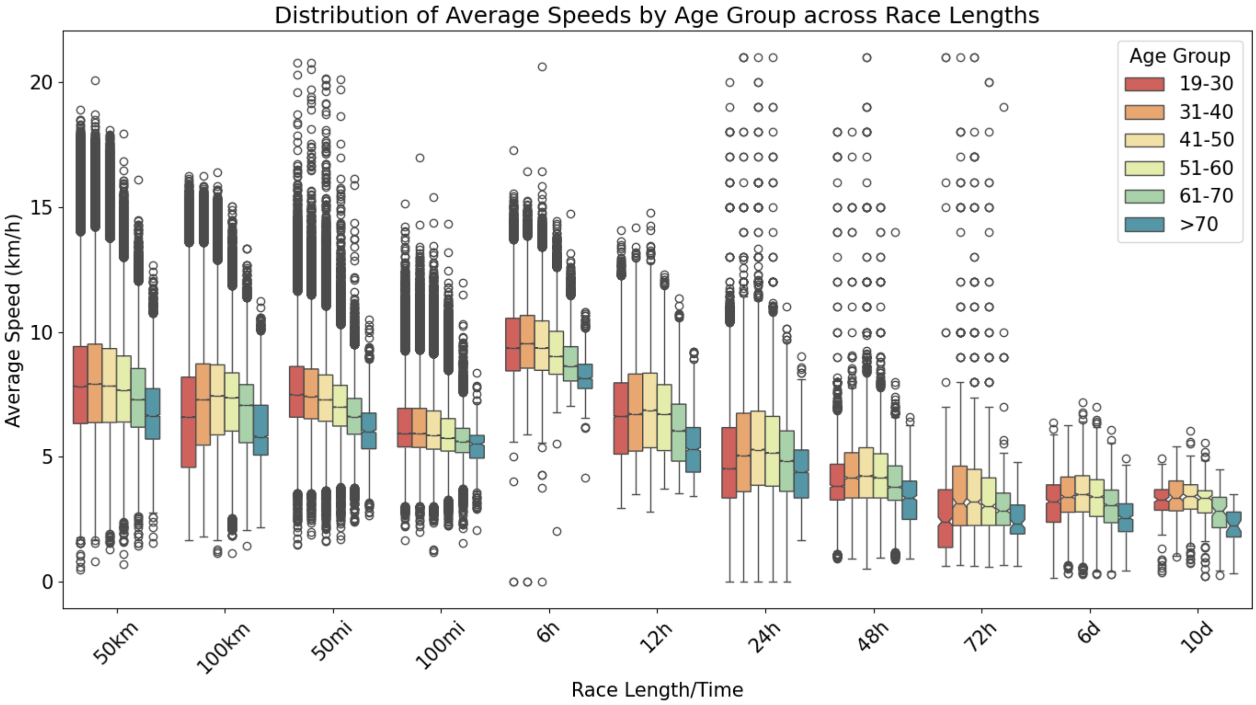

Figure 3: Average speed by age across the most popular type of race categories.

The results above can be explained by a mixture of health and experience from the athletes. For example, the 19-30 group may not achieve the highest speeds for some categories because of a lack of experience, while the 41-50 group might not perform well because of a general decline in health that comes with older age. The group above 70 years likely faces an even greater health decline, and finally, the 31-40 group exhibit the best balances of both experience and health. The reason why the younger groups perform better than the 41-50 group for fixed lengths could be because fixed length races are comparatively shorter, benefiting young athletes who have more energy, while the opposite happens for timed races as the 41-50 year group have more experience in their ability to navigate these longer races.

While knowing that male athletes between the age of 31 and 40 perform comparatively better than their peers is useful, it does not provide enough information to predict the performance of a new athlete, as there may be different attributes and also external factors that could potentially affect their average speed. In order to predict the performance, we will use an ordinary least squares (OLS) multiple regression model. The predictor variables are the athlete's age, gender, and country, as well as the year and season the event took place, the race category, and the event name itself. One-hot encoding was used on both the country and the event name in order to be able to run the regression on these categorical variables.

Unfortunately, there was low correlation between almost all of the predictor variables and the average speed, with the highest correlation being between the category and the performance ($r = -0.26$). Despite this, running the regression on all of the aforementioned predictors yielded the highest R^2 value of ≈ 0.149. This indicates that the model is not accurate in being able to predict a potential athlete's performance. There may be lurking variables that are unaccounted for. For example, the dataset contains no information about specific training regimens or data about the athletes themselves (especially since they were anonymised). Weather and types of terrain are also not accounted for. These factors have an impact on the performance of the athletes [12], as, naturally, worse weather and extreme terrain can make it difficult for athletes to perform better. Finally, athletes' psychological factors could also account for any possible deviances that may affect the model's predictability.

In an attempt to improve the regression model, we filtered the dataset by the top 5 most popular events. We define this to mean events with the most number of occurrences in the dataset across all years. The aim is to be able to account for geographical features across these 5 events, as they all occur in differing locations. It should be noted that these are the most popular events for the specific race categories that were filtered for as mentioned in the data processing section, not for all possible race categories available in the original dataset. As expected, once filtered the dataset, the average speed and race category were the most correlated variables ($r = -0.62$), as is evident in Figure 4. Longer races should, of course, lead to lower performance due to extended periods of physical action and a higher endurance requirement. The season was also closely correlated with the performance ($r = -0.60$). It makes sense that more extreme weather would negatively affect athlete performance, which is why the correlation coefficient is negative.

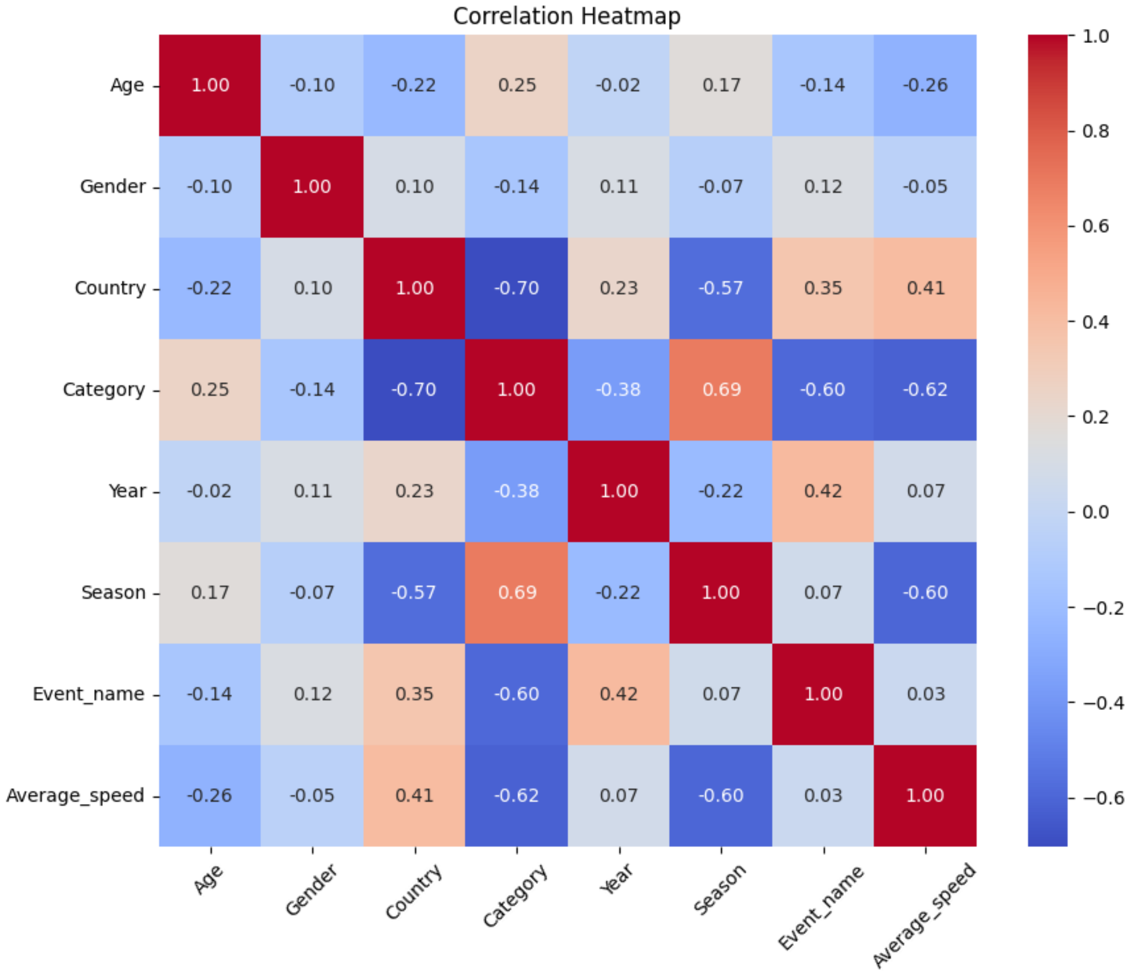

Figure 4: Correlation heatmap between all of the predictor variables and the average speed.

After re-running the regression model, we obtained an improved R^2 value of ≈ 0.498. Because the category and season were the most correlated, they had the highest β coefficients of -2.47 and -0.86 respectively, as can be seen in Table 1, suggesting that these are the most significant factors when it comes to evaluating the performance. Also, despite being smaller than the category, the gender had a negative β coefficient (-0.72), which suggests that males perform better than females, although slightly. This is in agreement with the previous analysis that concluded the same.

We interpret the improved R^2 value to be partially due to being able to categorise each of the top 5 events in terms of topography. The original dataset did not include any information with regards to this, and so, by filtering by the 5 most frequently occurring races, we are able to categorise the topography of each event's route. These events fell into 3 main "extremes": relatively flat, slight ascent, and steep — these are the differences in elevation athletes would encounter. Despite the race category being the same, a change in topography would significantly affect the performance, as steeper climbs would definitely cause a decrease in pace. Another factor that was implicitly accounted for by filtering were the terrain types. It should be noted that, for these 5 tracks, the terrain type was road (asphalt) [13, 14, 15]. Had the terrain type been different, it would have likely affected the performance, as running on, say, dirt roads may prove to be more difficult than on asphalt – even for the same race category. This may have been another unaccounted factor that led to the initially low R^2 value.

Table 1: OLS Regression Results

Dep. Variable:	Average_speed	**R-squared:**	0.498
Model:	OLS	**Adj. R-squared:**	0.498
Method:	Least Squares	**F-statistic:**	3.495e+04
Date:	Mon, 01 Apr 2024	**Prob (F-statistic):**	0.00
Time:	15:26:18	**Log-Likelihood:**	-4.0494e+05
No. Observations:	211752	**AIC:**	8.099e+05

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	138.0309	1.295	106.590	0.000	135.493	140.569
Age	-0.0223	0.000	-71.984	0.000	-0.023	-0.022
Gender	-0.7220	0.009	-80.341	0.000	-0.740	-0.704
Country	-0.0073	0.000	-55.940	0.000	-0.008	-0.007
Season	-0.8637	0.006	-153.370	0.000	-0.875	-0.853
Year	-0.0606	0.001	-94.451	0.000	-0.062	-0.059
Category	-2.4682	0.013	-197.187	0.000	-2.493	-2.444

To delve deeper into other possible explanatory factors that could enhance our regression model, we examined the impact of athlete recurrence—defined as the frequency of participation in races—on average speed, and assessed if this influence was uniform across all types of ultra-marathon events. As illustrated in Figure 5, there is a discernible relationship between an athlete's recurring appearances in races and their average speed across most event categories. It is a plausible assumption that increased participation correlates with gains in experience and training, which in turn could elevate the athlete's sustained speed. Notably, the 48-hour and 72-hour events diverge from this pattern. In these extended-duration events, the effect of recurrence on average speed appears to be less pronounced. This suggests that, for ultra-endurance races of such lengths, other variables—potentially including the strategies for rest and sleep—might exert a more significant influence on performance outcomes. All in all, the more experienced an athlete was, the more likely they were to have a higher performance.

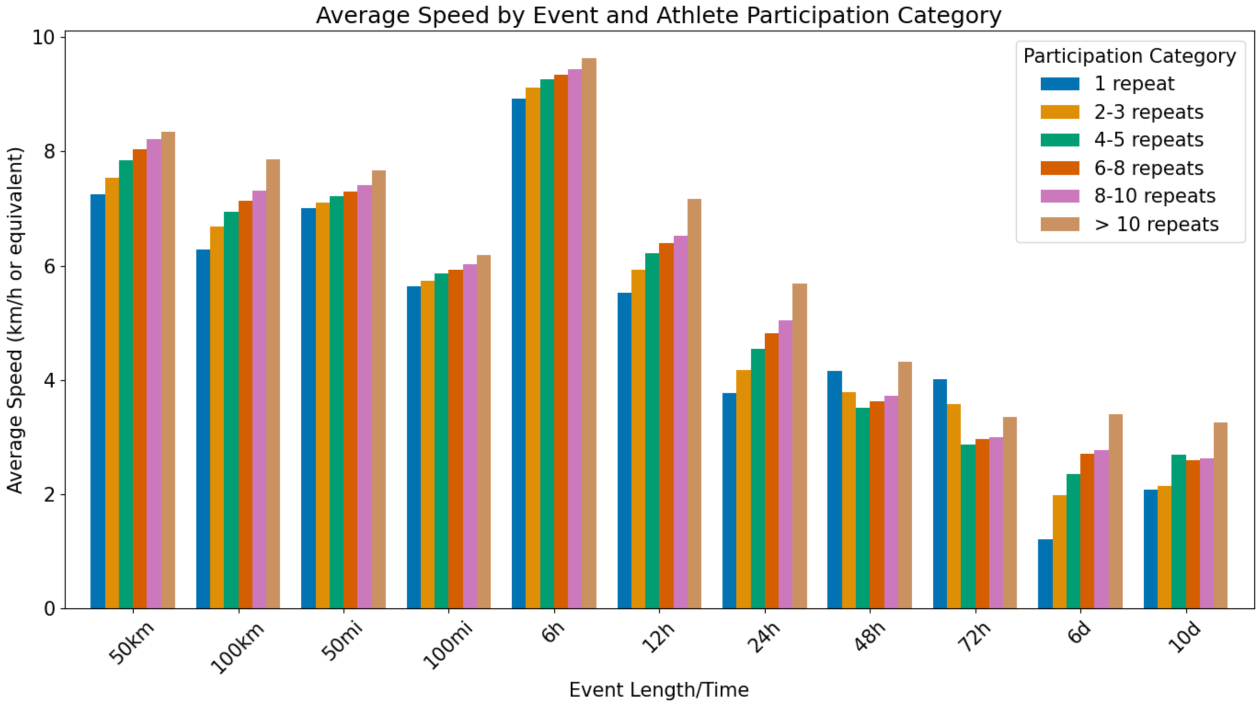

Figure 5: Average speed by athlete experience, quantified by the number of times they have participated in that race type, across the most popular race categories.

5 Discussion and conclusions

Summary of findings Our analysis revealed that, in recent years, Europe has been overtaken by North America as the most popular continent of origin for ultramarathon participants. Also, we discovered that males have a higher average speed than females. We then found that the most successful athletes fall within the 31-40 age group, followed by the 19-30 group for fixed-length categories, and the 41-50 group for fixed-time categories. Furthermore, we identified the factors influencing performance in ultramarathons and developed a model to predict future performance based on those factors. Initially, we were not able to accurately predict performance, with our model achieving a low R^2 value of 0.149 . We attributed this low value to mainly not accounting for topography and terrain type.

We then refined this prediction model, incorporating the aforementioned factors (along with season and year), and further improved its accuracy by focusing solely on the 5 most popular events. This approach yielded a satisfactory R^2 of 0.498. In support of previous analysis, the gender yielded a negative β coefficient, suggesting that males were overall faster than females, although the coefficient itself was small compared to other predictor variables. Additionally, in preparation for future studies, we explored how an athlete's experience impacts average speed, finding that an athlete's performance improves with increased participation in a specific type of event, indicating that experience was likely an important factor with regards to average speed.

Evaluation of own work: strengths and limitations After a significant amount of work, we believe that we have thoroughly explored and analysed trends in ultra-marathon running. Our visualisations delve into the characteristics of the athletes, and our findings are consistent with previous research in the field. The extensive dataset further supported our regression model; even after filtering for the top 5 events, we still obtained a large number of observations. This, in turn, lent credibility to our regression model and our significant R^2 of 0.498.

The primary limitation of our work lies in the reliability of our regression model. While using the top 5 events provides a solid prediction, the dataset's lack of information on terrain type and topography diminishes the model's utility for unknown events with different geographical features. Another limitation is the analysis of only the most common types of events. This excludes consideration of events like the Comrades Marathon, an event with a non-standard length category (88 km) that, nonetheless, attracts the greatest number of participants.

Comparison with any other related work Our findings are consistent with previous literature in the field. The best performing athletes belonged to the 31-40 age group according to a report by Knechtle and Nikolaidis [16] and the fact that men were slightly faster than women is also analysed. However, while previous work in the field analysed the impact of single characteristics (age and gender) on the performance, as far as we know, no other paper built a regression model to predict the performance also taking into consideration the country of origin and season. This type of analysis was previously done only for marathons so we believe that our research provides extensive insight for more extreme versions.

Improvements and extensions If given the means and time, we would extend the report by finding the best performing regions and looking for common characteristics. Additionally, we aim to further improve our regression model by incorporating additional datasets that include a wider variety of races, weather, track type information (such as the terrain of the route), and even specific athletes' training regimens—all of which can potentially refine our model. These improvements could lead to more accurate predictions, thereby assisting future athletes in adjusting their training regimens based on these insights.

References

[1] Wikipedia contributors. *Ultramarathon — Wikipedia, The Free Encyclopedia*. [Online; accessed 27-March-2024]. 2024. URL: https://en.wikipedia.org/w/index.php?title=Ultramarathon&oldid=1215477978.

[2] P. Matthews. *Historical Dictionary of Track and Field*. G - Reference,Information and Interdisciplinary Subjects Series. Scarecrow Press, Incorporated, 2012. ISBN: 9780810867819. URL: https://books.google.co.uk/books?id=dQFHe9RwE0wC.

[3] Severin Carrell and Scotland correspondent. "World's oldest race celebrates 500 years". In: *The Guardian* (June 2008). URL: https://www.theguardian.com/uk/2008/jun/23/scotland (visited on 03/27/2024).

[4] Jessica Murray. *'These races are epic': why ultrarunning is soaring in popularity*. the Guardian, June 2021. URL: https://www.theguardian.com/lifeandstyle/2021/jun/14/these-races-are-epic-why-ultrarunning-is-soaring-in-popularity.

[5] Volker Scheer et al. "Age-related participation and performance trends of children and adolescents in ultramarathon running". In: *Research in Sports Medicine* 28.4 (2020). PMID: 32573266, pp. 507–517. DOI: 10.1080/15438627.2020.1781124. URL: https://doi.org/10.1080/15438627.2020.1781124.

[6] Martin D. Hoffman and Kevin Fogard. "Demographic Characteristics of 161-km Ultramarathon Runners". In: *Research in Sports Medicine* (). PMID: 22242737. DOI: 10.1080/15438627.2012.634707. eprint: https://doi.org/10.1080/15438627.2012.634707. URL: https://doi.org/10.1080/15438627.2012.634707.

[7] David Valero and Elias Villiger. *The big dataset of ultra-marathon running*. www.kaggle.com, 2023. URL: https://www.kaggle.com/datasets/aiaiaidavid/the-big-dataset-of-ultra-marathon-running/data (visited on 03/27/2024).

[8] *David | Contributor*. www.kaggle.com. URL: https://www.kaggle.com/aiaiaidavid (visited on 03/27/2024).

[9] Creative Commons. *Creative Commons — CC0 1.0 Universal*. Creativecommons.org, 2019. URL: https://creativecommons.org/publicdomain/zero/1.0/.

[10] Sean McAlister. *How fast was Eliud Kipchoge's 2022 Berlin Marathon world record?* Olympics.com, Sept. 2022. URL: https://olympics.com/en/news/how-fast-was-eliud-kipchoge-world-record.

[11] C. T. M. Davies and M. W. Thompson. "Aerobic performance of female marathon and male ultramarathon athletes". In: *European Journal of Applied Physiology and Occupational Physiology* 41 (Aug. 1979), pp. 233–245. DOI: 10.1007/bf00429740. (Visited on 03/29/2020).

[12] Nicolas Bouscaren, Guillaume Y. Millet, and Sebastien Racinais. "Heat Stress Challenges in Marathon vs. Ultra-Endurance Running". In: *Frontiers in Sports and Active Living* 1 (2019). ISSN: 2624-9367. DOI: 10.3389/fspor.2019.00059. URL: https://www.frontiersin.org/articles/10.3389/fspor.2019.00059.

[13] *100 KM – Bieler Lauftage – Les courses de Bienne*. 100km.ch. URL: https://100km.ch/en/100km/ (visited on 04/01/2024).

[14] *L'Ultra 100Km*. 100 km de millau, Dec. 2023. URL: https://100kmdemillau.com/epreuves/ultra-100km/ (visited on 04/01/2024).

[15] *Ultra Marathon*. Two Oceans Marathon. URL: https://www.twooceansmarathon.org.za/events/ultra-marathon/ (visited on 04/01/2024).

[16] Beat Knechtle and Pantelis Theodoros Nikolaidis. “The age of the best ultramarathon performance – the case of the “Comrades Marathon””. In: *Research in Sports Medicine* 25.2 (2017). PMID: 28114817, pp. 132–143. DOI: 10.1080/15438627.2017.1282357. eprint: https://doi.org/10.1080/15438627.2017.1282357. URL: https://doi.org/10.1080/15438627.2017.1282357.