UltraInsight: Analyzing Paces, Ages, and Trends in Ultramarathon Finishers

ULTRAINSIGHT ' 24

Claire Lueking
Applied Computer Science
University of Colorado Boulder
Boulder, CO USA
claire.lueking@colorado.edu

1 PROBLEM STATEMENT

The knowledge gained through the mining of this dataset could help ultramarathoners plan their races through pacing strategies. Additionally, this information could be used in recognizing which countries and athletes are at the top of the field based on metrics and race statistics. Thirdly, the information could be used to determine which countries are more popular for ultramarathons and how the trend of increasing or decreasing finishers could be related to ultramarathon participation.

Through analysis of the dataset I hope to find out what are the average paces sustained, per race distance, for ultramarathon finishers. Additionally, I hope to find what the typical ages are of ultramarathon finishers that finish in the top 20% of the field per race distance. Lastly, I hope to find out which year(s) had the most ultramarathon finishers with paces in the top 20% of the field per race distance.

2 LITERATURE SURVEY

Prior work done on the subject of ultramarathons includes a study of master's athletes being examined for peak age and performance trends via pace and other measures for 24 hour ultramarathons in a 13 year study [1]. Another study focused on successful finishers of ultramarathons by assessing them for performance in more than 2000 100 kilometer races over 59 years to find out running speed and finisher age trends throughout the years. This study was worldwide [2]. A third study was done that analyzed pacing strategies of male elite and age groups ultramarathon racers to find trends related to age and race distance [3].

3 PROPOSED WORK

The dataset will be reduced to the past ten years of racing included in the dataset, from 2012 to 2022 in order to simplify the dataset to make it more manageable to run analyses on. An age column will be added in order to describe the numeric age of the participants to make it easier for analysis. Timed races will be removed in favor of set distance races for simplified analysis. Many times, timed races do not mean they are the same time for everyone, especially if the format is a lapped race. In timed lapped races, sometimes there is not enough time to start and/or finish another lap, thus someone's time is less than the official race time. Additionally, timed lap races do not have easily measurable paces for the entire race distance. This is because many people do not run at a consistent pace or take breaks in between laps, thus messing with pace in a way continuous races do not. The distances of the races will be converted to kilometers if they are in miles in order to match the majority of races, which are in kilometers.

In terms of data cleaning, entries that have blank attribute cells will be removed. Information will be verified by manually looking up random samples of race results. If there are a large number of discrepancies, each race will be evaluated individually for accuracy in reporting, though the dataset is anticipated to be accurate based on preliminary searches already.

As for deriving data, I plan to use birth years and race years to calculate the age of an athlete for the specific race by subtracting the birth year from the race year. Additionally, I plan to create a pace column to calculate the pace per race entry in minutes per kilometer. This helps ensure that later data analysis has the paces it needs to calculate average paces per race distance as well as find the standard deviation per race distance when looking for outliers. For average pace per race distance, I plan to find the mean($\bar{x}$)

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N} \quad (1)$$

for the paces [4]. In order to find the top 20% of the field per race distance, I plan to use a loop that goes over the selected distances and finds the top 20% of the field via the top runners according to pace.

These measures will be useful in comparing athletes and finding out key points of information related to the attributes and their calculations. I also intend to use standard deviation (σ)

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}} \quad (2)$$

to find outliers within the data as well [4].

For the project design I plan to follow the milestones outlined in section 7 in order to code and analyze the dataset. The coding format will be organized according to the data preprocessing and cleaning needs, the order of questions that need to be answered, and residual work to make the results more clear before moving onto visualizations. For the report design, it will be organized in the ACM SIG format (as this paper is organized in it as well) and the specifications outlined by the project.

For evaluation, I plan to use the evaluation methods as outlined in section 5 of this paper in order to verify and cross reference information and results. These results will be compared to established results from the literature review, section 2, studies.

4 DATASET

The dataset is a collection over seven million race records of ultramarathon finishers from 1798 and 2022. Ultramarathons are any race distance over 42.195 kilometers, ranging from 50 kilometers to over 320 kilometers. The data came prom public websites with the race results. To make the dataset more private regarding athlete names, the names were removed and an athlete identification attribute was made. In addition to that attribute, the attributes of year of event, event date(s), event name, event distance, event number of finishers, athlete performance, athlete club, athlete country, athlete year of birth, athlete gender, athlete age category, and athlete average speed are included within the dataset. The dataset URL is here: https://www.kaggle.com/datasets/aiaiaidavid/the-big-dataset-of-ultra-marathon-running [5].

5 EVALUATION METHODS

To evaluate the findings, the results will be compared to the results of the studies mentioned in the literature review. To compare the paces, an absolute and percent difference formula will be used. This can verify if paces are approximately around where they should be for the top twenty percent of the field. An example of this would be finding the average pack for the 50km distance for 18-22 year-olds in the dataset and comparing it with the similar age range pace in the studies mentioned in section 2's literature review.

To compare the years or age ranges, we will look at how they overlap with the established studies. This can help verify if the approximate ages and years match up with the average ages of the ultramarathoners who compete and finish the races as well as the year(s) of racing for finding the top 20% of paces. An example of the age range verification would be calculating the age range that has the top 20% of paces for a certain distance, and comparing it with the previously mentioned studies' results. If there is some overlap, it is a good conclusion that the age range could be modified slightly. However, if there is not much overlap, the data might need to be looked at or recalculated again. An example of the year verification would be to look at who finished in the top 20% of paces in a particular year and compare that with the previous studies' results to see if there is overlap in agreement.

Additionally, the plausibility of the results will be cross referenced with outside race results and other studies if more verification is needed. All calculations done with equations, such as those that will be listed in the tools section, will be verified at least three times to make sure that the answer is stable and not changing each time.

6 TOOLS

A tool that will be used is Microsoft Excel for file reading and basic data cleaning. The class textbook will also be used for equations such as the ones for mean (μ), standard deviation (σ), and correlation (ρ) for statistics and comparison between attributes (labeled as Equations 1,2, and 3 above).

Additionally, the Python programming language within VS Code to make visualizations and calculate statistics. Python packages such as numpy, pandas, and matplotlib will be used to aid this process. Lastly, the equations from the book will be used in conjunction with coding to help guide aspects of the analysis.

7 MILESTONES

The data preprocessing will be done by March 30, 2024. As for analysis of the data, that will be done by April 10, 2024. The progress report will be done by the due date of April 15, 2024. For creating visualizations, those will be done by April 20, 2024. The final report will be done by April 30, 2024 and the presentation will be done by the due date of May 2, 2024.

7.1 Milestones Completed

The data preprocessing has been completed as of March 30, 2024. This ended up being the most challenging part of working with the dataset as there were multiple formatting errors with the dataset including the data not being numeric, but string-based, the athlete performance column having multiple string characters that were not easily removed as they had different formats. It was important to remove these characters as it was preventing the data from becoming numeric, something that was needed for the analysis section.

The dataset was successfully reduced to the past 10 years of results. The age category column was modified to be an age column due to the age category column holding the actual ages, just with prefixes to distinguish them in male and female categories. The age was verified by taking birth year and matching it to the race year to get the age of the athlete for that specific race.

The gender data was stored elsewhere in the dataframe, thus it was okay to remove the prefixes in the age column.

Timed races were removed in favor of set distance races for simplified analysis. The distances, if they were in miles, were converted to kilometers in order to match the majority of races, which were in kilometers. Additionally, a pace column was added to show the pace per race entry in minutes per kilometer.

In terms of data cleaning, entries that had blank attribute cells were successfully removed. Duplicates of entries were also removed. Information was verified by manually looking up random samples of race results. There were no large number of discrepancies in the race results, thus not necessitating individual evaluation of every race result. The data was very dirty so general cleaning took a lot of time as much of it had to be done by hand for certain columns such as event distance and athlete performance.

One section of the data analysis was completed by April 10, 2024. For average pace per race distance, I found the mean ($\overline{x}$) for the paces per distance. Additionally, I found the top 20% of the field according to pace per race distance. Lastly, I also found the standard deviation (σ) to find outliers within the pace per distance data as well. The data analysis proved to be very difficult and time consuming, especially with the challenging data cleaning process, so it took longer than anticipated to finish it. The progress report was completed by April 14, 2024.

7.2 Milestones To-Do

The visualizations will be done by April 20, 2024. The final report will be done by April 28, 2024 and the presentation will be completed by the project due date of May 2, 2024.

8 RESULTS SO FAR

For the average pace for race distance, there was one for each selected distance. The paces were rounded to two decimal points for simplicity. The average pace for 50 km races, it was 8.13 minutes per kilometer. For 80 km (50 mi) races, I found that the average pace was 8.87 minutes per kilometer. For 100 km races I found that the average pace was 9.13 minutes per kilometer. Lastly, the average pace for 160 km (100 mi) races was 10.45 minutes per kilometer.

These results were interesting because as the longer the race distance became, the longer it took to go a kilometer within the races. This makes sense as the longer the race is, the longer it takes to complete. Additionally, longer races tax the body more, and people can slow down as a result of their body getting more and more tired as the race goes on.

For the average pace (rounded to two decimal places for simplicity) of the top 20% of the field for each selected race distance, the results were interesting. For 50 km races, the average top 20% of the field's pace was 11.36 minutes per kilometer. For 80 km (50 mi) races, the average top 20% pace was 11.85 minutes per kilometer. For 100 km races, the average top 20% of the field's pace was 13.28 minutes per kilometer. Lastly, for 160 km (100 mi) races, the average top 20% of the field's pace was 13.51 minutes per kilometer.

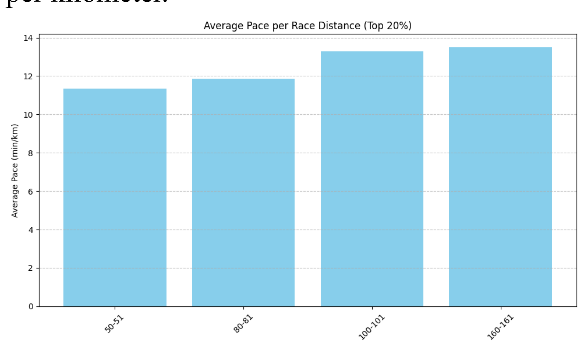

Figure 1. This figure shows the average pace in minutes per kilometer for selected race distances of 50 km, 80 km, 100 km, and 161 km for the top 20% of the field per race distance according to pace. Note: Many ultramarathons are not exactly measured thus a range of values within 1 kilometer was used for the bin of pace.

These results were notable because while they seem slow to road racing standards, these paces are very fast. An additional consideration is that many of these races take place on mountainous terrain where the terrain grade can typically be up to 40-plus% grade. Unfortunately, the percent grade data was not included in this dataset, making it impossible to make grade-adjusted-pace calculations for different races with different grades. This checks out with the top 20% of the field for each selected race distance being in the elite percentage of the entire field.

For the race distances, standard deviation was used to see if there were any outliers by pace within the dataset. These paces were also rounded to two decimal places. For 50 km races, paces less than 4.86 minutes per kilometer and greater than 11.29 minutes per kilometer were outliers.

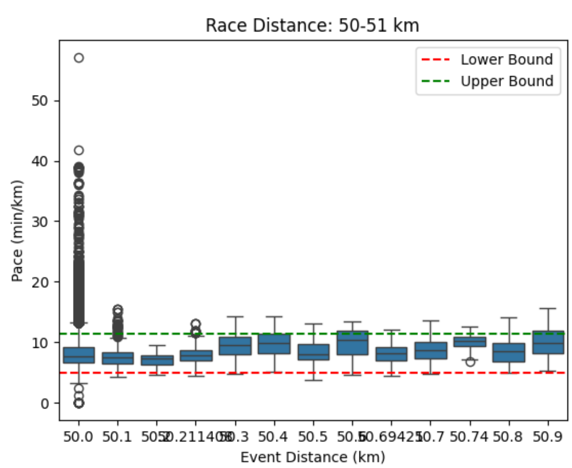

Figure 2. This figure shows box-and -whisker plots for 50 km races and outliers by pace in minutes per kilometer. The upper and lower bounds are the average bounds where outside of those bounds (above the upper bound and below the lower bound) there are typically outliers. Note: Many ultramarathons are not exactly measured thus a range of values within 1 kilometer was used for the bin of pace.

For 80 km (50 mi) races, paces less than 5.82 minutes per kilometer and greater than 11.86 minutes per kilometer were outliers.

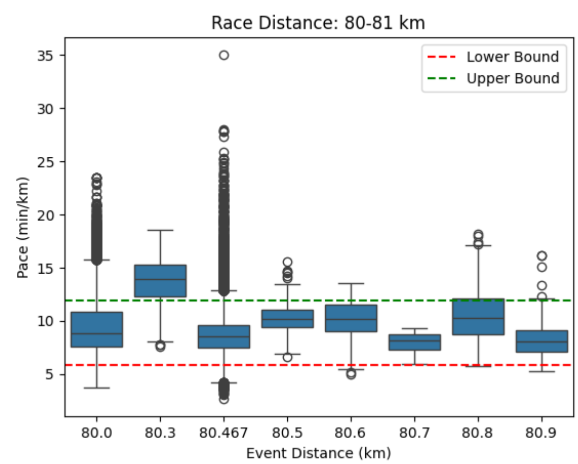

Figure 3. This figure shows box-and -whisker plots for 80 km races and outliers by pace in minutes per kilometer. The upper and lower bounds are the average bounds where outside of those bounds (above the upper bound and below the lower bound) there are typically outliers. Note: Many ultramarathons are not exactly measured thus a range of values within 1 kilometer was used for the bin of pace.

For 100 km races, paces less than 5.05 minutes per kilometer and greater than 12.90 minutes per kilometer were outliers.

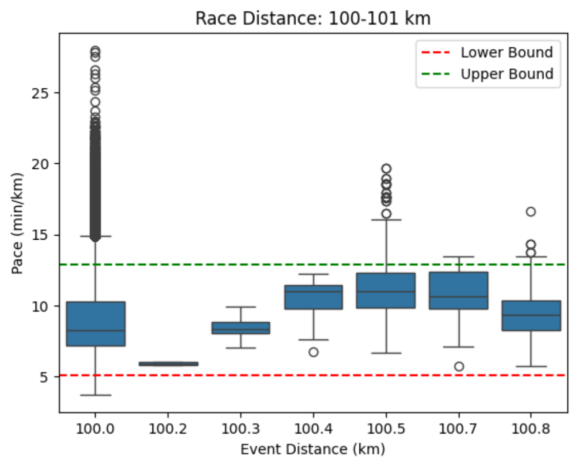

Figure 4. This figure shows box-and -whisker plots for 100 km races and outliers by pace in minutes per kilometer. The upper and lower bounds are the average bounds where outside of those bounds (above the upper bound and below the lower bound) there are typically outliers. Note: Many ultramarathons are not exactly measured thus a range of values within 1 kilometer was used for the bin of pace.

For 160 km (100 mi) races, paces less than 7.28 minutes per kilometer and greater than 13.59 minutes per kilometer were outliers.

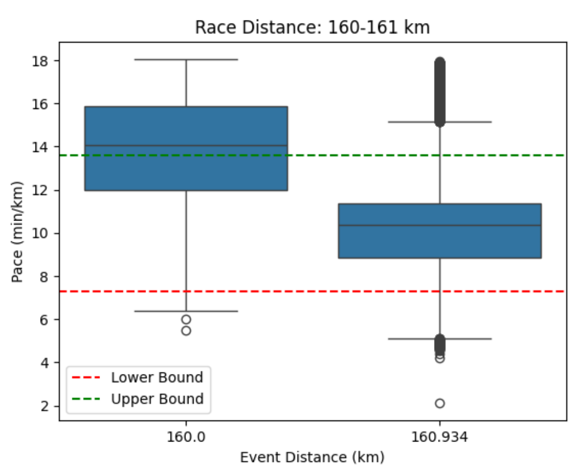

Figure 5. This figure shows box-and -whisker plots for 160 km races and outliers by pace in minutes per kilometer. The upper and lower bounds are the average bounds where outside of those bounds (above the upper bound and below the lower bound) there are typically outliers. Note: Many ultramarathons are not exactly measured thus a range of values within 1 kilometer was used for the bin of pace.

These times make sense because ultramarathon distances usually have slower paces than road races, but for the 50 km distance, the pace isn’t that much slower than a road marathon. For 161 km (100 mi) races, being slower than 13.59 minutes per kilometer can put someone in danger of not finishing the race within what are typically strict cutoff times. Of course, there are exceptions that can be seen within some of these plots, but generally the rule of thumb that above the upper bounds and below the lower bounds are outliers holds true. Something else interesting to note in summary, that there are more fluctuations within the 50 km distance than the 160 km distance because 50 km is a more popular distance than 160 km, and races are not usually measured exactly.

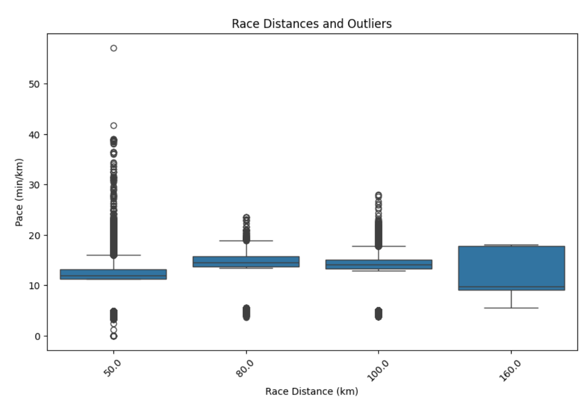

Figure 6. This graph shows a summary of race distances and their average paces as well as outliers per distance.

Overall, these results suggest that the analysis is accurate when it comes to common race paces of finishers and paces of those that finish in the top 20% of the field according to pace. Additionally, the results also suggest that there exist outliers within the dataset for being either faster or slower than the lower and upper bounds, but most of the data falls within those two bounds.

REFERENCES

[1] ZINGG, M., RÜST, C.A., LEPERS, R., ROSEMANN, T., AND KNECHTLE, B. 2013. Master runners dominate 24-H ultramarathons worldwide-A retrospective data analysis from 1998 to 2011 - extreme physiology & medicine. *BioMed Central*. https://extremephysiolmed.biomedcentral.com/articles/10.1186/2046-7648-2-21.

[2] STÖHR, A., NIKOLAIDIS, P.T., VILLIGER, E., ET AL. 2021. An Analysis of Participation and Performance of 2067 100-km Ultra-Marathons Worldwide. *National Library of Medicine*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7825131/.

[3] KNECHTLE, B., ROSEMANN, T., ZINGG, M.A., STIEFEL, M., AND RÜST, C.A. 2015. Pacing strategy in male elite and age group 100 km ultra-marathoners. *National Library of Medicine*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4376307/.

[4] HAN, J., KAMBER, M., AND PEI, J. 2011. *Data Mining: Concepts and techniques 3rd edition*. Elsevier Science, San Diego, CA, USA.

[5] DAVID. 2023. The big dataset of ultra-marathon Running. *Kaggle*. https://www.kaggle.com/datasets/aiaiaidavid/the-big-dataset-of-ultra-marathon-running.