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

ABSTRACT

Four questions came into mind when looking into this ultramarathon dataset. What are the average paces sustained over ultramarathons for finishers? What are the average paces of the top 20% of the field in each race distance? What year(s) had the most finishers with paces in the top 20% of the field per race distance? What outliers are there with pace for finishing certain distance ultramarathons?

A summary of the results includes the average pace per race distance differing by distance and it generally gets slower as the race distance increases. The average pace per race distance for the top 20% of the field per race distance. The paces also differ slightly for each distance, but generally as the race distance increases, the pace becomes a little slower.

As for the years with the most finishers broken down by race distance, the years were different except for 50 km races and 160 km (100 mi) races, which had the same year. This was most likely a coincidence, but could have been due to extraneous factors happening in the world that were not included in the dataset.

As for outliers, most outliers were above the upper bounds and a few were below the lower bounds broken down by race distance. The exceptions to the general rule that outliers exist above the upper bounds and below the lower bounds were mostly in the 50 km race section. The most likely reason for this is that because 50 km races are so popular, there is a lot of variation in pace because more people with different goals do them. Another reason could be that time cutoffs for these races could differ from the longer races by being more forgiving with cutoffs, thus increasing variations in pace.

CCS CONCEPTS

• Data Structures and Algorithms • Database Systems • Software Development Fundamentals • Programming Languages • Data Management • Computer Systems Organization • Human-Centered Computing • Computing Methodologies

KEYWORDS

Python, Pandas, Data Processing, Data Cleaning, CSV Files, String Manipulation, Numeric Conversion, Data Filtering, DataFrame Operations, Time Calculations, Concatenation, Column Insertion, Data Readability, Data Analysis, Documentation

1 Introduction

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 help make predications for races by sports broadcasters and news outlets. Fourthly, it could help 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 the hope is to find out what are the average paces sustained, per race distance, for ultramarathon finishers. Additionally, the hope is to find out what paces the top athletes in each race distance field are running to be in the top 20%. Thirdly, the hope is to find out which year(s) had the most ultramarathon finishers. Lastly, the hope is to find information regarding outliers and how those play into finisher statistics and average paces for those racing in ultramarathons.

2 Literature Survey

Prior work done on the subject of ultramarathons included a few studies. There was a study of master's athletes that were examined for peak age and performance trends. This study was conducted by looking at masters' athletes running 24-hour ultramarathons over the course of a 13-year study. Measurements were taken by looking at their paces over the races and age grouping categories. The results showed that both female and male age group runners improved running speed and those greater than 40 years old had the fastest running paces [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]. The results showed that there are an increasing number of ultramarathon events, thus helping to drive the increase in number of male and female participants. Additionally, they also found that those over 40 years old increased their performances more than others in younger age groups.

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]. The results showed that for age group athletes that running speed was decreased over the course of the races, with the exception of the last segment of some races where 40 plus year old's increased their speeds. Additionally, it was found that altitude had no effect on performance for the top ten runners with age group athletes

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, the plan is 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, the plan is 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, the plan is 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, the plan is 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. The plan is to also 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 the plan is 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 clearer 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, the plan is 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 Main Techniques Applied

Many techniques were used to work with this dataset. Some techniques used fall into the categories of data cleaning and preprocessing as well as data warehousing.

5.1 Data Cleaning and Preprocessing

For data cleaning, removal techniques were used to remove duplicate entries within the data as well as entries that had blank sections in them. This type of cleaning was because it seemed inauthentic to try to extrapolate data that was not there, especially when dealing with something as personal is individual race performance finish times, paces, and more. Timed races were also removed due to many factors that were mentioned in section 4 that would make it hard to keep the data from being skewed in analysis.

As for data preprocessing, the dataset was reduced from over seven million entries to a little less than three million entries after being cleaned for simplified analysis. An existing age category column was repurposed as a flat age column as the categories were just the age of the athlete with a prefix of "M" for man and "W" for woman in front of it. The prefixes were removed and this was made possible by the gender and sex data being stored as another variable already within the dataset.

5.2 Data Warehousing

Data warehousing techniques were used in a few places. Aggregate statistics such as calculating averages and counts were used in pace calculations, especially within the top 20% of the field per race distance pace calculations. Additionally, there was aggregation by time period in the case of finding the years that had the most ultramarathoners per distance. This is a common technique used to help identify trends over time within data warehousing. Lastly, trends and patterns were identified within pace categories per race distance. This is also a common technique used in data warehousing and data analysis.

6 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.

7 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.

8 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 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.

8.1 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 ($\bar{x}$) 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. The visualizations were done by April 14, 2024. The final report was finished on April 15, 2024. The presentation was completed by April 18, 2024.

9 Key Results

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.

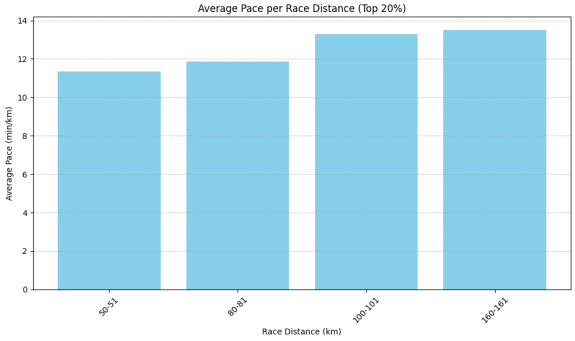

Figure 1. This figure shows the average pace in minutes per kilometer for selected race distances of 50 km, 80 km, 100 km, and 161 km for the top 20% of the field per race distance according to pace. Note: Many ultramarathons are not exactly measured thus a range of values within 1 kilometer was used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

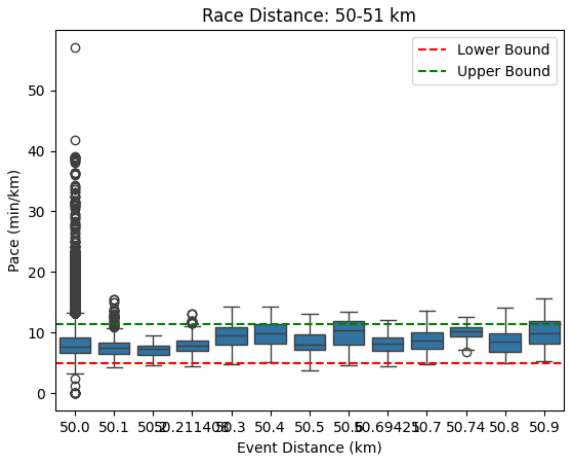

Figure 2. This figure shows box-and -whisker plots for 50 km races and outliers by pace in minutes per kilometer. The upper and lower bounds are the average bounds where outside of those bounds (above the upper bound and below the lower bound) there are typically outliers. Note: Many ultramarathons are 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.

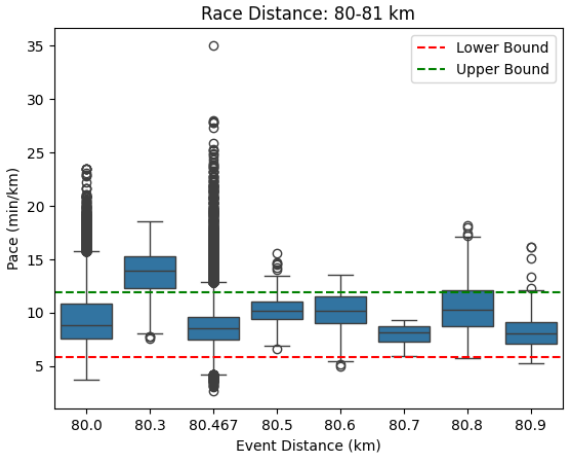

Figure 3. This figure shows box-and -whisker plots for 80 km races and outliers by pace in minutes per kilometer. The upper and lower bounds are the average bounds where outside of those bounds (above the upper bound and below the lower bound) there are typically outliers. Note: Many ultramarathons are 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.

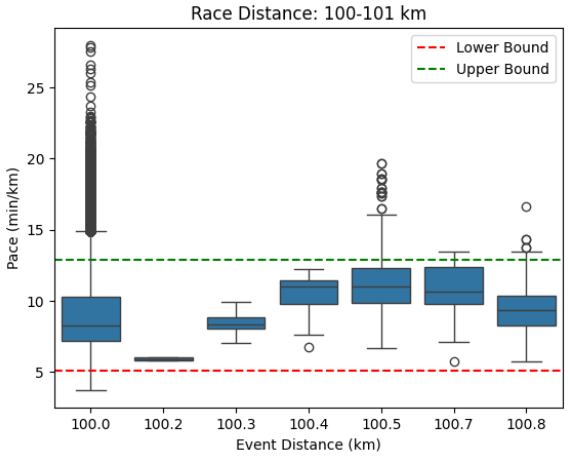

Figure 4. This figure shows box-and -whisker plots for 100 km races and outliers by pace in minutes per kilometer. The upper and lower bounds are the average bounds where outside

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.

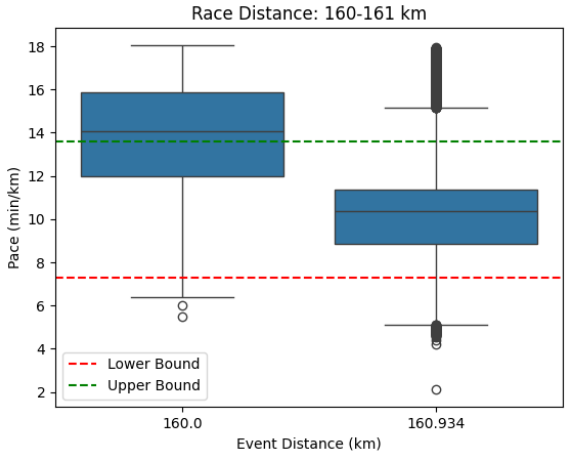

Figure 5. This figure shows box-and -whisker plots for 160 km races and outliers by pace in minutes per kilometer. The upper and lower bounds are the average bounds where outside of those bounds (above the upper bound and below the lower bound) there are typically outliers. Note: Many ultramarathons 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.

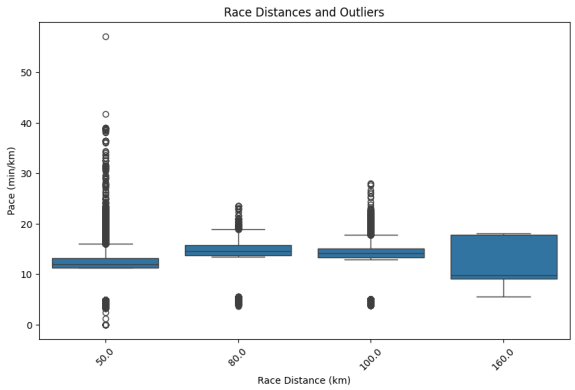

Figure 6. This graph shows a summary of race distances and their average paces as well as outliers per distance.

This summary figure shows the most important parts of the race distances, their average paces, and outliers. For the 50 km distance, there are more outliers than the 160 km race because the 50 km distance is so popular that the different races have different time standards and thus more variation in average pace.

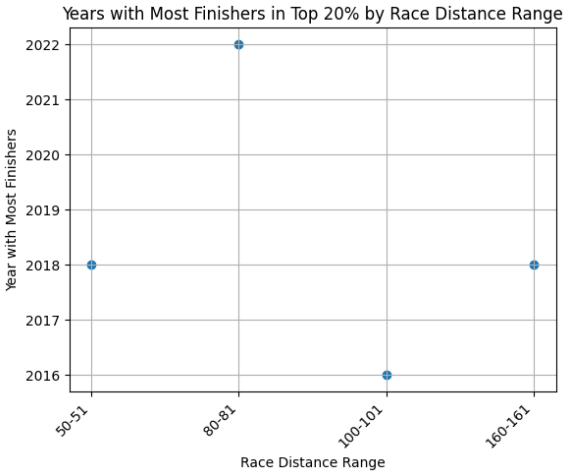

Figure 7. This figure shows a scatterplot of the years with the most finishers broken down by race distance. Note: Many ultramarathons are not exactly measured thus a range of values within 1 kilometer was used for the bin of pace.

This figure illustrates that different race distances have different years where there are the most finishers. The interesting result is that 50 km and 160 km races have the same year with the most finishers. This is most likely a coincidence barring other extraneous information that was not mentioned in the dataset.

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.

10 Applications

The knowledge gained from the results can be applied in many different ways. One of those ways is that the average pace data can be used to help someone plan their race according to the average pace for the race distance they are going to run. Average pace tells someone a lot about how fast they need to run across the race in order to finish. From the average pace starting point, they could strategize how to cover different types of terrain and what are realistic paces for covering that terrain. These could all factor into meeting the average pace. From there, they could customize this to their goal pace.

The knowledge gained from recognizing the top 20% of the field per race distance’s average pace is that runners, especially those close to or in the top 20% of their race field could use this average pace to plan their personal race pace. Of course, they would still also have to factor in percent grade of the terrain just the same as those basing their race based on average pace for the whole field per the specific race distance. Additionally, this information could give those in or near the top 20% an idea of what pace they need to train at or work up to before the race to be competitive during the actual race.

A third application of the results is that knowing the upper and lower bounds of the standard deviation is also helpful in planning race paces, but also in predictions for finish times according to sports broadcasters. There are races that put on a broadcast of the race and even predictions of races about who will place where. This information about the upper and lower bounds could help broadcasters to make realistic predictions about who will finish and when by making sure the paces fall in between these two bounds unless they know of people who are exceptions to those ranges. Additionally, standard deviations can help runners predict if their training paces are sound to focus on finishing a race, regardless of performance. Many ultramarathons have strict cutoff times throughout the race, and knowing if their paces fall outside the average range could help them adjust their pace as need during the race so that they meet the checkpoints under certain cutoff times.

12 Visualizations

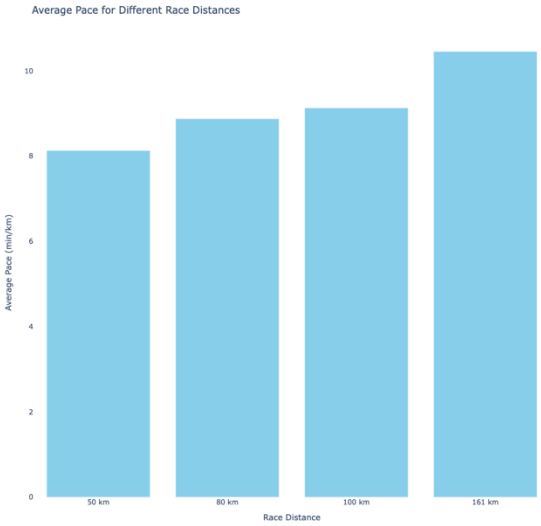

Figure 8. This is an interactive bar graph visualization showing the average pace for different race distances. The interactive visualization is linked here. Note: Many ultramarathons are not exactly measured thus a range of values within 1 kilometer was used for the bin of pace.

This visualization summarizes the mean pace per each race distance in according to the entire fields of each race distance. The visualization shows the exact pace per each race distance grouping when you hover over the bars. This is important to illustrate as it shows that each different distance has an average pace, and that those generally increase as the distance increases. This alludes to the idea that the longer the race distance, the more the pace has to be adjusted for the body to keep going for as long as it can at the pace chosen by the runner.

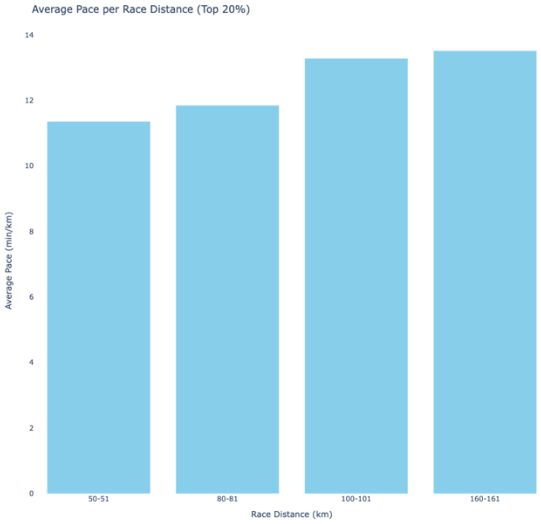

Figure 9. This is an interactive bar graph visualization showing the average pace for the top 20% of the field for different race distances. The interactive visualization is linked here. Note: Many ultramarathons are not exactly measured thus a range of values within 1 kilometer was used for the bin of pace.

This visualization summarizes the average pace per race distance for the top 20% of the field in each race distance. In this visual, hovering over the bars gives the exact average pace per each race distance grouping. This is important to illustrate as this shows what pace it takes, on average to be in the top 20% of the field in each of the selected race distances. Additionally, it can help athletes with training to try to stay in or break into the top 20% of the field.

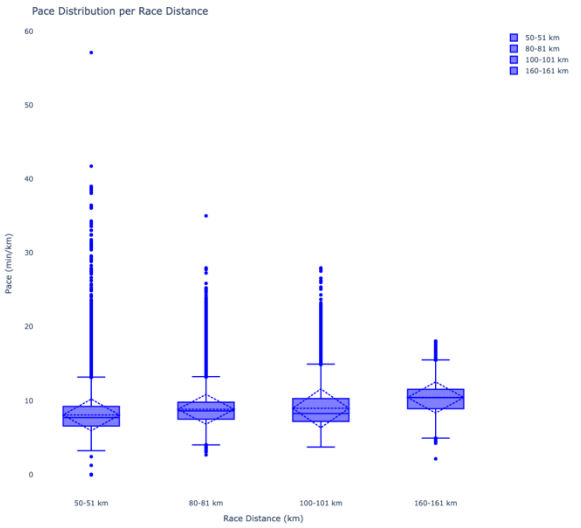

Figure 10. This is an interactive boxplot visualization showing the average pace for different race distances and their outliers. The interactive visualization is linked here. Note: Many ultramarathons are not exactly measured thus a range of values within 1 kilometer was used for the bin of pace.

This visualization shows various measures related to average pace per each race distance in addition to outliers for each distance. Included is the average, median, quartiles 1 and 3, and outliers for pace per each boxplot. This is important to illustrate as it can tell athletes if their training paces are feasible to finish these race distances. Additionally, it shows that the longer the race gets, the less variation in average paces, most likely due to people having to adjust for cutoff times that get stricter as distances get longer.

12 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.