What Makes a Marathon Runner Fast? Key Predictors of Finishing Times*

How age, gender, nationality, and race experience shape the journey to faster finishes.

Sophia Brothers

November 27, 2024

This paper develops a generalized additive model for marathon finishing times using factors such as age, gender, country, pace, and race experience. The analysis identifies key trends, including that younger runners, male participants, Kenyan athletes, and individuals with moderate NYRR race experience achieve the fastest time. These findings provide information into the characteristics that influence marathon performance, contributing to a deeper understanding of endurance running. This research provides important knowledge for athletes, coaches, and sports organizations, helping them design better training strategies and improve marathon performance.

Table of contents

*Code and data are available at: https://github.com/sophiabrothers1/marathonfinishers.

1 Introduction

Marathon running has grown exponentially in popularity over the past few decades, attracting participants from diverse backgrounds and skill levels to races of varying distances and difficulty levels (Hillier 2024). While the surge in participants reflects the global appeal of marathon events, it also brings to light a significant challenge: predicting marathon performance. Understanding the factors that influence marathon finishing times is important not only for individual runners but also for race organizers, coaches, and sports scientists seeking to optimize training regimens, race strategies, and overall athlete performance.

The estimand of this study is the causal effect of demographic and experience-based variables on marathon finishing times. Specifically, this research aims to quantify the influence of age, gender, country of origin, and cumulative race experience on a runner's overall marathon finishing time. By identifying and modeling these relationships, we estimate how variations in these predictors impact the outcome of interest: marathon finishing time.

The factors influencing marathon times are numerous and complex, including age, gender, previous race experience, health, and training practices. Existing research has highlighted the impact of certain demographic variables, such as age and gender, on marathon performance (Cheuvront et al. 2005), but there remains a gap in the understanding of how these and other factors quantitatively influence finishing times. For instance, how does a runner's previous race experience impact their finishing time, and are there interactions between these factors that could provide more context to performance predictions?

This paper seeks to address this gap by developing a model to predict marathon finishing times based on several key variables: age, gender, country of origin, and previous race experience. By using a generalized additive model (GAM), this study analyzes the non-linear relationships between these variables and their effect on marathon performance. The primary goal of the study is to quantify how these demographic and experience-based factors contribute to a runner's overall marathon finishing time, providing both individual runners and race organizers with important context into what influences race outcomes.

The results of this analysis have real-world implications for both individual runners and broader performance strategies. For runners, understanding the factors that most significantly impact their performance allows for more informed goal setting and training focus. For race organizers, identifying key predictors of performance could help in tailoring race-day experiences, such as pacing strategies, nutrition plans, and event management. Moreover, this information could help coaches better design training programs and guide runners on the areas that need improvement.

The paper is structured as follows: Section 2 discusses the data and data cleaning process, followed by the analysis of key variables and relationships in Section 3. Section 4 then presents the results from the GAM, followed by a discussion of the implications of these findings in Section 5. Section 5 also concludes with suggestions for future research and practical applications of the model.

2 Data

2.1 Overview

This study utilizes NYC marathon finishers' data that was sourced through the Data is Plural newsletter (Hovde 2024). This data provides detailed information on select runner demographics, race results, and other performance-related metrics. The dataset includes variables such as runner identity, demographic information (age, gender, city, country), and race performance data (finishing times, pace, rankings). These variables together form a profile of marathon participants, enabling the analysis of how different factors contribute to marathon outcomes.

We used R (R Core Team 2023) for data cleaning and analysis, using packages such as **`tidyverse`** for data manipulation and visualization (Wickham et al. 2019), **`here`** for file path management (Müller 2023), **`lubridate`** for date handling (Spinu, Grolemund, and Wickham 2023), **`arrow`** for data storage and processing (Richardson et al. 2023), **`testthat`** for unit testing (Wickham 2023b), **`readr`** for reading structured data (Wickham, Hester, et al. 2023), **`dplyr`** for data manipulation (Wickham, François, et al. 2023), **`stringr`** for string operations (Wickham 2023a), **`ggplot2`** for creating visualizations (Wickham, Chang, et al. 2023), **`caret`** for modeling (Kuhn 2023), **`kableExtra`** for creating tables (Zhu 2023), and **`modelsummary`** for summarizing models and results (Arel-Bundock 2023).

2.2 Measurement

The measurement process refers to how we go from real-world phenomena—such as a runner's time in a race, their age, or their gender—to numerical entries in a dataset. Each entry represents a distinct runner's marathon performance and is recorded in the dataset.

Overall Time (`overall_time`): This is the time it takes for a runner to complete the marathon, measured in hours, minutes, and seconds. It is recorded by the timing system used during the race, which uses a chip timer. Timing mats are located at the start, every 5K, halfway (13.1 miles), mile 20, and the finish. The data in the dataset records this time in a standardized format (HH:MM:SS), which is then converted into numerical values (seconds) for modeling purposes.

Age (`age`): This is the age of the runner at the time of the marathon. It is typically self-reported at the time of registration and stored in the race's database. The dataset uses the reported age as a direct entry for each runner.

Gender (`gender`): The gender of the runner is recorded at the time of registration and stored as a categorical variable (e.g., male, female, other).

Races Count (`races_count`): This variable reflects the number of marathons or races a runner has participated in. This variable is typically gathered from historical race records with New York Road Runners, the organization that hosts the NYC Marathon.

IAAF Category (`iaaf_category`): This is a country code that indicates the nationality of the runner, as classified by the International Association of Athletics Federations (IAAF). It is recorded during registration based on the runner's stated nationality or residency. The data is stored as a standardized three-letter country code (e.g., USA, CAN, GBR) and serves as a categorical variable.

2.3 Data Cleaning

The raw marathon finisher data underwent a several cleaning steps to ensure it was accurate, consistent, and ready for analysis. These steps included renaming columns for clarity, converting variables to appropriate data types, handling missing values by assigning default placeholders, and transforming time-based variables (e.g., overall time and pace) into numerical formats for modeling purposes. This resulted in the creation of two columns, `overall_time_seconds` and `pace_seconds`, which converted the overall time and pace data into seconds. The cleaned dataset was then saved as a Parquet file for efficient storage and further analysis.

2.4 Outcome Variables

The outcome variable is **`overall_time_seconds`**. This is the primary dependent variable that the model is designed to predict. It represents the total time (in seconds) a runner takes to finish the marathon. The model aims to understand the factors that influence this time, which is a direct measure of performance in the race. Figure 1 displays the actual finish times, measured in seconds, with the majority of finishers completing the race within 3.5 hours to 4.5 hours.

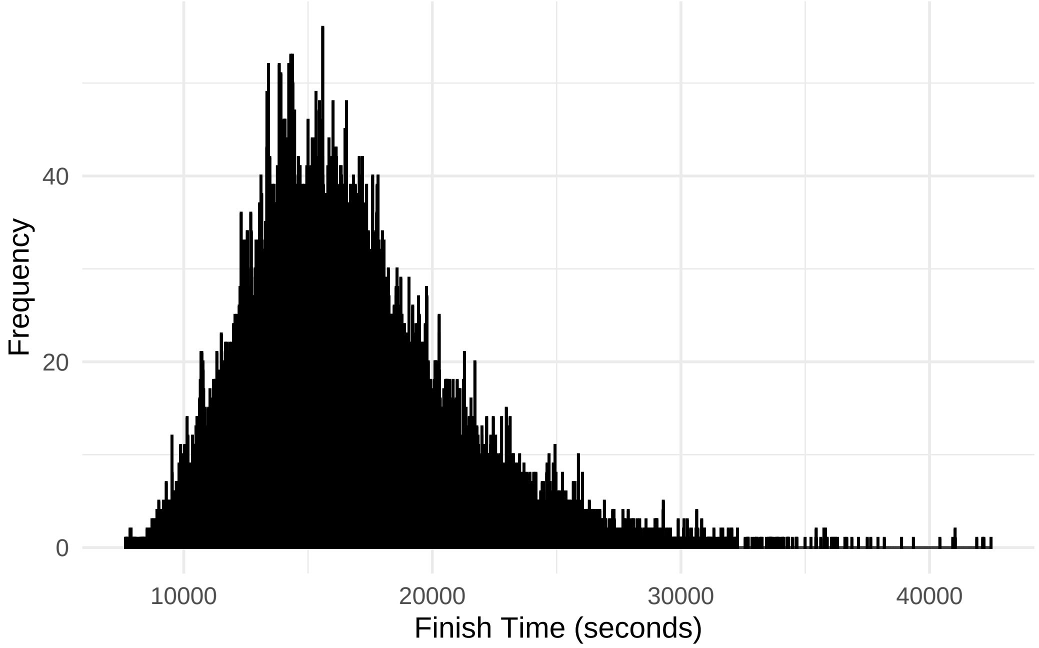

Figure 1: Distribution of marathon finish times, highlighting the concentration around the 4-hour mark. The skew towards longer finish times reflects the diverse abilities of participants.

2.5 Predictor Variables

The **predictor variables** (or independent variables) are the factors believed to influence the overall finishing time:

1. **Age (`age`)**: The runner's age is a key factor influencing marathon performance. The age value is an important variable as it directly relates to the physical capabilities of the runner, with younger runners often performing better, though not always in a linear fashion. Ages 25-40 are the most popular ages represented in the NYC marathon, as depicted in Figure 2.

2. **Gender (`gender`)**: The gender of the runner is included as a categorical variable (male/female/other). This is a key variable in the dataset because gender has been found to correlate with marathon performance, where physiological differences typically result in different average finishing times. In Figure 4, we see that the vast majority of runners identify as either Male or Female, with a marginal amount identifying as Other. There are more Male runners than Female runners.
3. **Races Count (`races_count`)**: This variable represents the number of races a runner has participated in, which can be an indicator of experience but is also correlated with age. It is an indirect measurement of experience, with the assumption that more experienced runners tend to perform better. Figure 5 shows that the majority of runners were completing their first race. Interestingly, there is a minor spike at 10, likely because of NYRR's 9+1 program that guarantees entree to the marathon (the 10th race).
4. **IAAF Category (`iaaf_category`)**: The IAAF category represents the country of the runner. In Figure 3 we see that the USA is the most represented country, which makes sense given that the race is located in New York City. Italy and France are the next most represented countries.

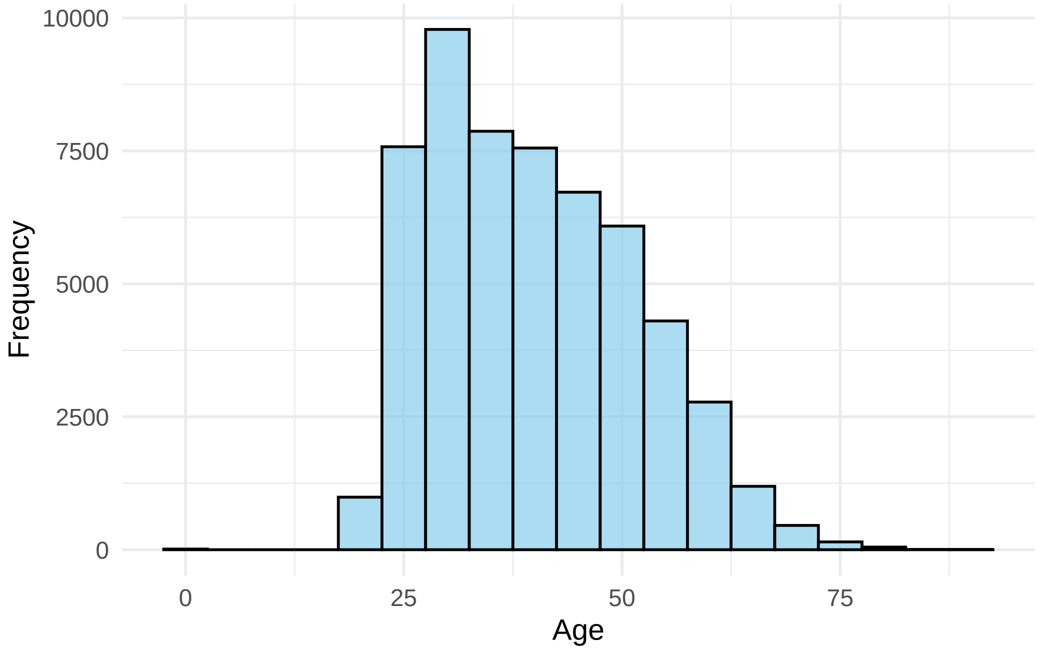

Figure 2: Distribution of runners' ages, showing a wide range of participation with a peak in the 30-40 age group, a common range for competitive runners.

These predictors were selected based on both domain knowledge of marathon running and the availability of data in the dataset. The model examines how these factors collectively

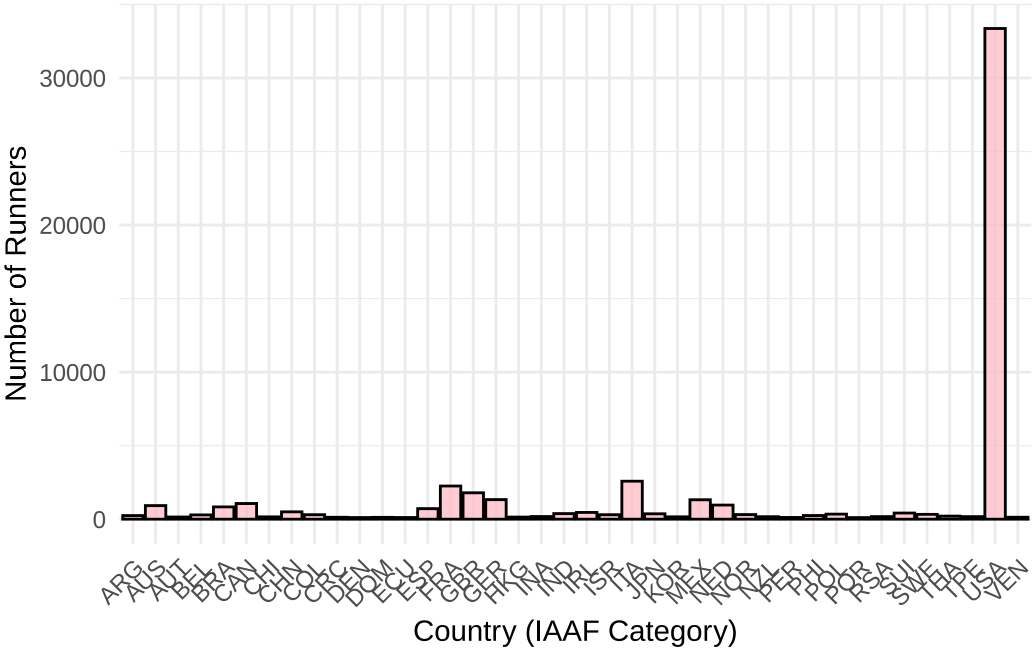

Figure 3: Distribution of marathon participants by country, limited to those with more than 100 runners. The United States has the highest number of participants, followed by several other countries with a smaller but significant representation.

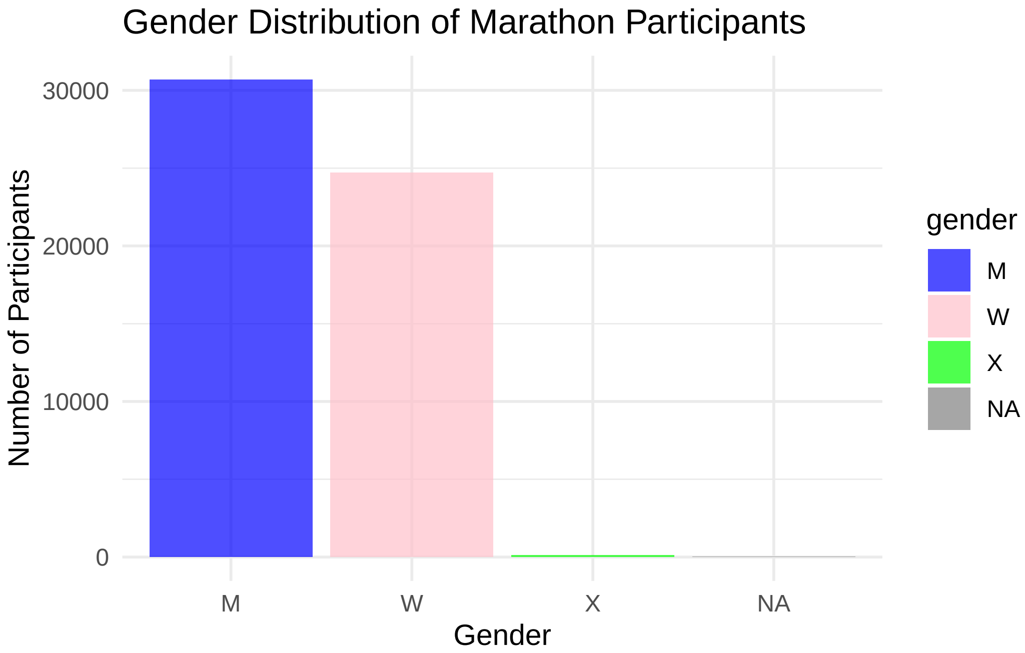

Figure 4: Gender distribution among marathon participants, showing a moderate predominance of male runners.

contribute to overall race performance, helping to provide information into which variables most strongly influence marathon finishing times.

3 Model

3.1 Model Set-Up

The goal of this analysis is to build a model using R (R Core Team 2023) to predict a runner's overall marathon finishing time in seconds (`overall_time_seconds`) based on several predictor variables. The outcome variable, `overall_time_seconds`, is continuous and represents the total time in seconds taken by a runner to complete the marathon. The GAM allows for capturing potential non-linear relationships between the predictors and the outcome.

The model is specified as follows:

$$overall\_time\_seconds_i = \beta_0 + s(age_i) + \beta_1 \cdot gender_i + s(race\_count_i) + \beta_2 \cdot iaaf\_category + \epsilon_i$$

Where:

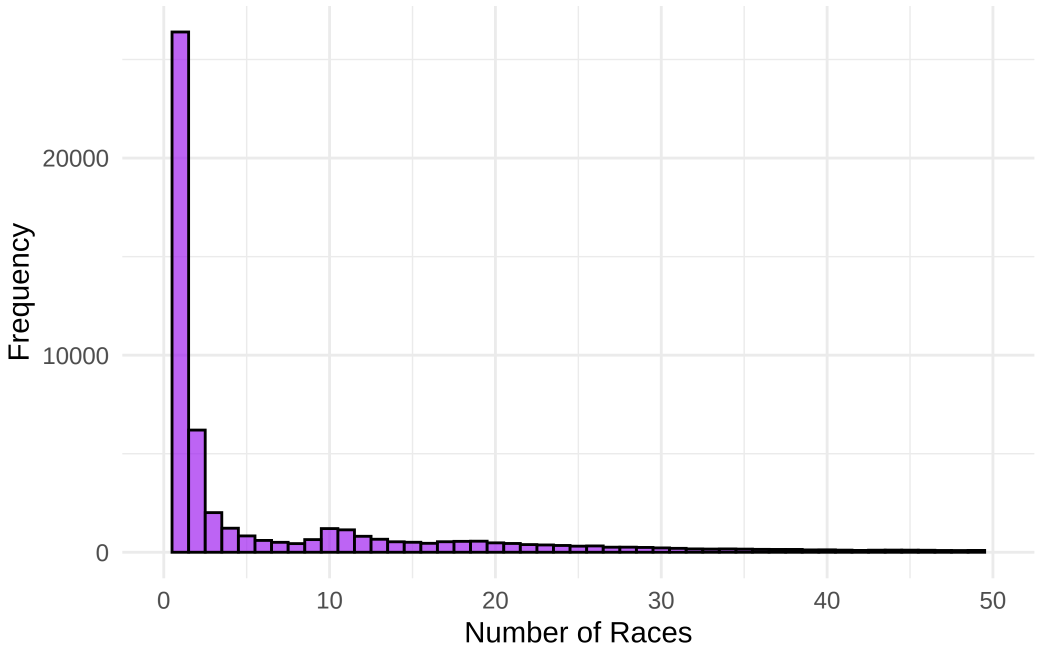

Figure 5: Distribution of the number of races completed by runners, indicating that most participants are relatively inexperienced, with a smaller number of highly experienced runners. Plot scaled from 0 to 50 races completed.

- $overall\_time\_seconds_i$ is the marathon finishing time in seconds for the i-th runner.
- $s(age_i)$ is a smooth function of the runner's age at the time of the race.
- $gender_i$ is the runner's gender (treated as a categorical variable).
- $s(race\_count_i)$ is a smooth function of the total number of NYRR races the runner has participated in.
- $iaaf\_category_i$ is the country the runner is from under IAAF standards.
- ϵ_i is the error term, assumed to be normally distributed with a mean of 0 and constant variance $\epsilon_i \ \sim N(0, \sigma^2)$

This GAM model allows for smooth, non-linear effects of continuous predictors (**`age`** and **`races_count`**) while keeping the categorical predictors (**`gender`** and **`iaaf_category`**) in a linear framework.

3.2 Model Justification

The GAM was chosen for this analysis due to its ability to capture non-linear relationships between predictors and the outcome variable, which is essential when dealing with continuous predictors like age and race count. Unlike traditional linear models, which assume a constant relationship between the predictors and the response variable, GAM allows for smooth, non-linear functions to model complex patterns in the data. Previous studies suggest that factors such as age and running experience do not have a simple linear relationship with marathon performance (Connick, Beckman, and Tweedy 2015). For example, the effect of age on performance might not be constant across all age groups, and the relationship between the number of races a runner has participated in and their finishing time could also exhibit non-linear trends. By using smooth functions (denoted as $s()$ in the model) for continuous predictors like age and race count, the GAM offers a more flexible approach that can better model these complex relationships without imposing a strict parametric form.

Moreover, while GAMs allow for these non-linear effects, they still retain interpretability. The model provides a clear visualization of how predictors like age and race count influence marathon times, with the smooth terms making it easy to understand the underlying trends. This interpretability is important in contexts where understanding the relationship between predictors and the outcome is as important as predictive accuracy.

The inclusion of gender and IAAF category as categorical variables further enhances the model's flexibility, as these variables likely have a linear relationship with marathon performance but differ across groups. By treating these as factors in the model, we account for the variations in marathon finishing times due to these categorical influences while keeping the overall model simple and interpretable.

Alternative models, such as linear regression and random forests, were also considered. The linear regression model, while simple and easy to interpret, assumes that the relationship between the predictors and the outcome is linear. Given the potential non-linear nature of the relationships in this dataset, this assumption may be too restrictive. Additionally, linear regression does not accommodate interaction effects between predictors without explicitly adding interaction terms, which could lead to an incomplete model. While linear regression is often a good starting point, it was deemed insufficient for capturing the complexities in the data in this case.

Random forests were also looked at as an alternative. They are highly flexible, non-parametric models that can handle non-linear relationships and interactions between predictors. Random forests are powerful in terms of predictive accuracy and can capture complex patterns in the data. However, they are harder to interpret compared to GAMs, as the results are more difficult to break down into understandable relationships between predictors and the outcome variable. Interpretability is a key priority in this analysis, as we are not only interested in prediction but also in understanding the underlying factors that influence marathon performance. For this reason, despite random forests' strong predictive power, they were not selected.

3.3 Assumptions and Limitations

The Generalized Additive Model (GAM) makes several key assumptions that must be considered when interpreting the results. One assumption is that the relationships between the continuous predictors (such as age and race count) and the outcome variable (overall marathon time) can be adequately captured by smooth functions. While GAMs are flexible and allow for non-linear relationships, they still rely on the assumption that these smooth functions are appropriate representations of the underlying trends. If the true relationships are more complex or if additional predictors are needed, the model may not fully capture the data's complexity.

Additionally, the model assumes that the residuals (the differences between the predicted and actual marathon times) are independent and identically distributed, a standard assumption for most regression models. This ensures that statistical tests, such as hypothesis tests for the significance of predictors, are valid. The model also assumes homoscedasticity, meaning that the variance of the residuals should remain constant across the range of predictors. If the variance of the residuals changes with different levels of the predictors, it could indicate that the model is misspecified.

One limitation of the GAM is that it does not account for interaction effects between predictors. For instance, the combined effect of age and race count on marathon performance could be more complex than the sum of their individual effects. A more sophisticated model could capture these interactions, potentially leading to a better understanding of how different factors together influence marathon finishing times. Another limitation is that the model assumes all relevant predictors are included. Environmental factors such as weather conditions on race day, course difficulty, and the runner's health status are not considered, even though these could significantly affect finishing times.

Outliers and extreme values in the data may also impact the model, as they can disproportionately influence the estimates of the smooth functions. While the GAM is more robust to outliers than traditional linear models, the presence of extreme values could still lead to biased results. Furthermore, the model assumes that the relationships between predictors and the outcome are consistent across the entire dataset, which may not always hold true if there are subgroups of runners with distinctly different characteristics or behaviors.

Finally, the GAM model may not generalize well to other populations, particularly if the sample is not representative of broader runner demographics. Differences in training, race conditions, or runner health that are not captured in the model could limit its applicability to other contexts. Therefore, while the GAM provides important context for marathon performance, its assumptions and limitations should be carefully considered when applying the model's findings to different situations or populations.

3.4 Model Validation

To validate the performance of the model, we employed the following techniques:

1. **Out-of-Sample Testing**: The data was randomly split into training and test sets, with 80% used for training the model and 20% reserved for testing. This allows us to assess how well the model performs on data that was not used during training.
2. **Root Mean Squared Error (RMSE)**: The RMSE was calculated on the test set to evaluate the model's predictive accuracy. In Table 1 we see that the RMSE is 3383.13. The RMSE provides an estimate of the average error between the predicts and actual values of the outcome variable, meaning that the average error between the model and actual time is 3383.13 seconds (56.38 minutes). A lower RMSE indicates better predictive performance.

Extended versions of these tables can be found in Table 5 and Table 6.

4 Results

These statistics indicate that, on average, participants completed the marathon in approximately 4 hours and 31 minutes and 48 seconds (271.8 minutes). The average participant age was 39.89 years, and the average pace was 10.37 minutes per mile. 55,524 people completed the New York City Marathon.

Figure 6 compares the predicted and actual marathon times across different age groups. The plot shows a linear relationship between predicted and actual times, with the reference red

Table 1: Summary of the marathon finishing time model, which includes age, gender, and race count. The table presents the model coefficients along with their standard errors.

	(1)
GenderMale	13898.563
	(1694.547)
GenderFemale	1963.963
	(29.654)
GenderOther	1601.891
	(311.826)
Num.Obs.	55512
R2	0.166
AIC	1060138.9
BIC	1061745.5
RMSE	3383.13

Table 2: Summary statistics of the marathon data, including the average finish time, average age, and total number of participants analyzed.

AverageFinishTime	AverageAge	AveragePace	TotalParticipants
272	40	10	55524

dashed line representing perfect predictions. The data shows that older age groups tend to have higher finishing times, but the model performs consistently across age categories.

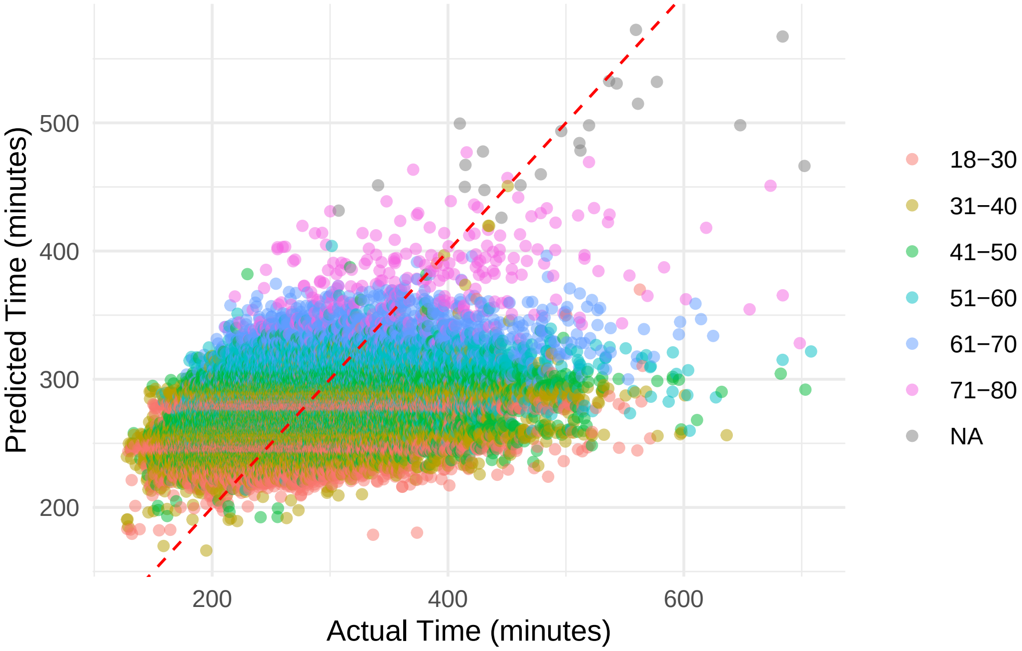

Figure 6: How well the predicted finishing times compare with the actual times across different age groups, with a reference line indicating perfect predictions.

The distribution of predicted marathon finishing times by gender was visualized using a boxplot in Figure 7. This analysis shows that the predicted times for males is lower than that for females, suggesting that the model may predict faster times for males on average. The distribution of predicted times for males tends to be slightly more spread out. The median predicted time for males is lower than that for females.

The boxplot shown in Figure 8 demonstrates the distribution of predicted marathon times across different age groups. As expected, older participants tend to have slower predicted finishing times, with the younger two age groups (18-30 and 31-40) showing the fastest predicted times on average.

Table 4 and Table 3 compare the actual best-performing groups with their corresponding predicted best groups across four categories: age, gender, number of races completed, and IAAF category. A category must have at least 25 entries/predictions to be considered.

1. **Age**

   - The actual and predicted best groups for Age were both 28, with closely aligned mean times: 15569.09 seconds (actual) and 15543.88 seconds (predicted).

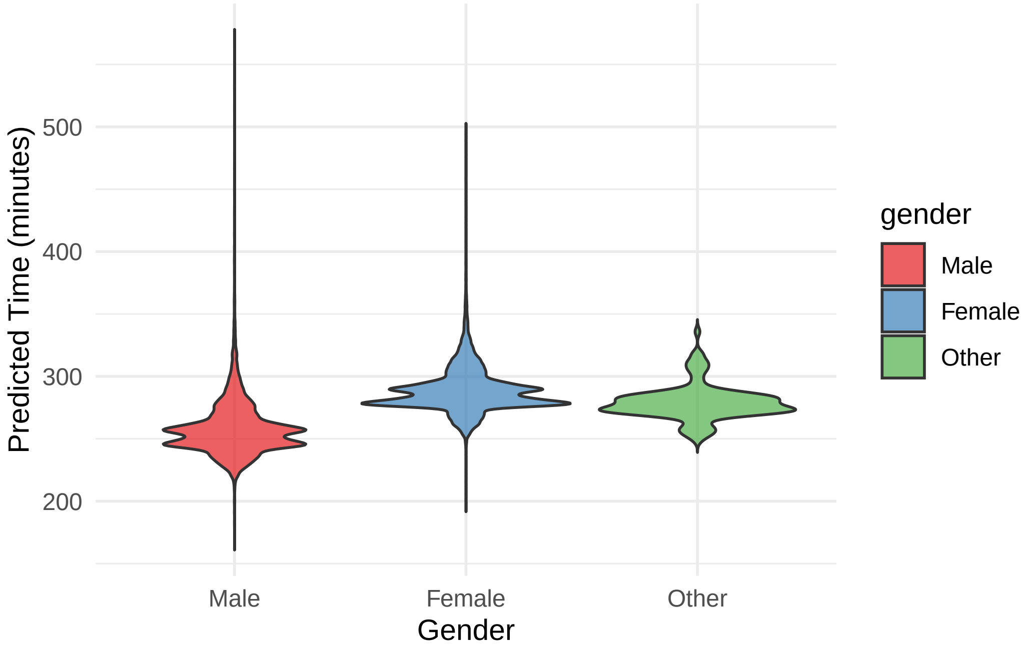

Figure 7: Distribution of predicted marathon times for different genders, highlighting the variability in finishing times.

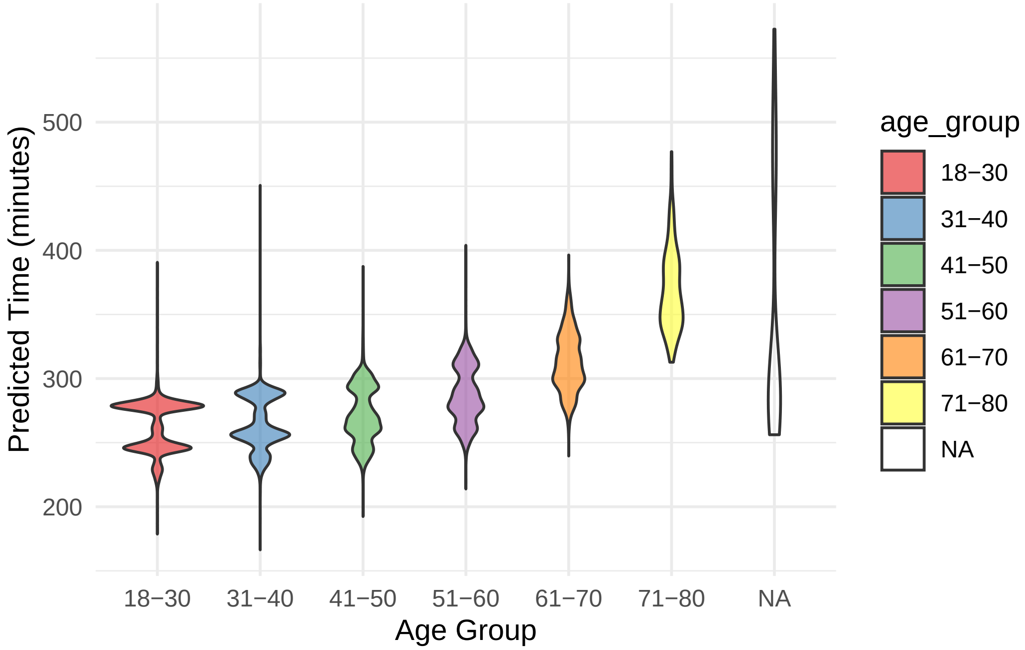

Figure 8: Distribution and density of predicted marathon times for various age groups

Table 3: This table shows the best group (age, gender, race count, IAAF category) based on actual marathon times.

Category	Actual Best Group	Actual Mean Time (seconds)
Age	28	15569.09
Gender	Male	15486.41
Races Count	5	15792.92
IAAF Category	KEN	12157.34

Table 4: This table shows the best group (age, gender, race count, IAAF category) based on predicted marathon times.

Category	Predicted Best Group	Predicted Mean Time (seconds)
Age	28	15543.88
Gender	Male	15486.41
Races Count	1	16067.05
IAAF Category	KEN	12157.34

2. **Gender**

   - The actual and predicted best groups for gender were both Male, with closely aligned mean times: 15,413.65 seconds (actual) and 15,423.57 seconds (predicted).

3. **Races Count**

   - **Actual Best Group**: Runners who participated in 5 NYRR races had the fastest mean actual time of 15,779.89 seconds, with a participant count of 818.

   - **Predicted Best Group**: The model predicted runners with 1 races as the best-performing group, with a slightly slower mean time of 16067.05 seconds.

4. **IAAF Category**

   - Both the actual and predicted best-performing group was KEN (Kenya), with mean times of 12,026.07 seconds (actual) and 12157.34 seconds (predicted).

5 Discussion

5.1 Predicting Marathon Finishing Times Using Demographic Data

This study aimed to develop a model for predicting marathon finishing times based on demographic factors such as age and gender. By analyzing the relationship between these variables and actual marathon performance, we can gain important knowledge into the factors influencing marathon results and how well demographic information can predict performance. The model appears to capture key trends, such as the general increase in finishing times with age, as well as differences between genders, with males tending to have faster predicted times than females. These findings are consistent with previous literature suggesting that younger participants and males generally perform better in marathons.

5.2 Age and Gender as Significant Predictors

One of the key findings of this study is the role of age and gender in predicting marathon finishing times. The model demonstrated that age is a significant predictor, with older participants generally having slower predicted times. This trend aligns with what is commonly observed in long-distance running, where younger athletes often perform better due to physical factors such as active muscle mass and maximal oxygen consumption (Connick, Beckman, and Tweedy 2015). The model also displayed a noticeable difference between genders, with males typically predicted to finish faster than females. This trend is consistent with general physiological differences between men and women, as distance running is determined largely by aerobic capacity and muscular strength, both of which men possess to larger degrees than women (Cheuvront et al. 2005).

5.3 Kenyan Runners Dominate in Results

Kenyan runners emerged as the fastest group, aligning with their global reputation in long-distance running. This dominance highlights the interplay of genetic, cultural, and training factors that contribute to exceptional performance. This aligns with previous studies which showed there are several factors that allow Kenyan distance runners to excel. This includes a genetic predisposition, high maximal oxygen intake, high hemoglobin, metabolic efficiency, favorable skeletal-muscle-fiber composition and oxidative enzyme profile, diet, altitude training, and motivation (Wilber and Pitsiladis 2012).

5.4 Limitations: Demographic Focus, Dataset Bias, and Generalizability

While the model provided important information into the relationship between demographic factors and marathon performance, there are several limitations to consider. First, the model relies solely on demographic data (age, gender, and country), excluding other potentially influential factors such as training history, diet, and health conditions, which can significantly impact marathon performance. Additionally, the model assumes that demographic factors alone are sufficient to predict marathon times, but in reality, performance is likely influenced by a complex interplay of physical, psychological, and environmental factors that are not captured in the data.

Another limitation is the potential bias in the dataset. If the dataset used for training the model is not representative of the broader marathon population, the model's predictions could be skewed, leading to less accurate results for participants outside the dataset's scope. Specifically, the dataset for this study is focused on participants in the NYC marathon, which could introduce biases related to geography, socio-economic status, and access to resources. For instance, the dataset may overrepresent runners from the NYC metro area, which is known for having a relatively high socio-economic status and access to professional training resources. As a result, the model may not generalize well to runners from more diverse or underserved backgrounds, who may face different challenges and have different training experiences that impact their performance. The marathon also requires a hefty fee in order to participate, further exacerbating socioeconomic bias.

Additionally, the model assumes that demographic factors, such as age, gender, and nationality, have consistent and independent effects on marathon performance. However, in practice, these factors might interact in complex ways. For example, older runners from certain countries may have access to better healthcare and training resources than younger runners from other regions, leading to differences in their performance outcomes that the model does not account for. The interplay of these demographic variables may cause certain groups to be over- or underrepresented in the predictions, resulting in biased or inaccurate estimates for certain populations.

Finally, the model's ability to predict finishing times may decrease for extreme cases, such as elite runners or those with unusual age profiles (e.g., very young or very old participants). These runners often possess unique characteristics, such as rigorous training regimens or physical conditions that are not captured in the dataset, making their performance less predictable based on demographic data alone. As a result, the model may not be suitable for predicting performance for these outliers, limiting its usefulness for certain subgroups within the marathon community.

5.5 Next Steps: Enhance Predictive Accuracy and Expand Model Scope

Future research should look at several avenues for improving the model. First, expanding the dataset to include additional variables such as training habits, health conditions, and previous race performance could provide a better view of the factors influencing marathon times. Incorporating these variables would likely enhance the model's accuracy and offer more personalized predictions.

While the current model uses a GAM, which allows for non-linear relationships between predictors and the outcome, marathon times may be influenced by more complex interactions that could be better captured by other modeling techniques like random forests or gradient boosting.

Another area for future research could involve looking at the impact of environmental factors, such as weather conditions and course difficulty, on marathon performance. These factors could be incorporated into the model to help provide more accurate predictions for races held under different conditions. Additionally, accounting for these external variables would contribute to a more holistic understanding of what drives marathon performance, particularly in varied race settings.

6 Appendix

6.1 Data Cleaning Notes

The cleaning process began with renaming all columns to standardized and intuitive names, improving readability and ensuring consistency with analysis conventions. Next, data types were adjusted for various variables to align with their intended use. Numerical fields, such as age, overall place, and race count, were converted to integers, while time-based variables like overall time and pace were transformed from string formats (e.g., HH:MM:SS or MM:SS) into total seconds for accurate modeling and analysis. Categorical variables, such as gender and country codes, were encoded as factors to enable statistical comparisons.

Missing values were addressed by assigning meaningful placeholders to relevant fields, such as substituting "Unknown" for missing first names, cities, and states, or "N/A" for missing IAAF categories. To facilitate numerical modeling, overall time was split and converted into total seconds, while pace was converted into seconds per mile, allowing for consistent performance comparisons.

Finally, the cleaned dataset was exported in Parquet format using the `arrow` library, chosen for its efficiency in storage and processing. These steps collectively ensured that the dataset was reliable, consistent, and optimized for in-depth analysis.

6.2 Idealized Methodology

What is the population, frame, and sample?

- **Population**: The population for this study would be all marathon runners globally, including recreational, amateur, and elite athletes who participate in marathons. This includes runners who participate in a wide variety of events, from local races to prestigious international marathons.

- **Frame**: The frame refers to the specific set of marathon runners who will be considered in the study. Ideally, this would include individuals who have participated in a marathon within the past year, providing up-to-date performance data. The frame could be obtained from event organizers, race registries, or online race databases that track marathon results, such as the *Marathon Handbook* or *World Athletics.*
- **Sample**: The sample would be a subset of marathon runners from the frame. To make this sample more manageable, it could be stratified by factors such as age, gender, race distance (standard marathon vs. ultra), or region. The sample could include both runners who have already participated in multiple marathons as well as first-time marathoners, ensuring diversity in the data.

How is the sample recruited?

- **Recruitment through Race Registrations**: Runners who register for marathons could be invited to participate in the study. During the registration process, participants could opt-in to share their marathon data, including times, training habits, and performance metrics, with the study organizers. This can be done through a consent form or a checkbox during the registration process.
- **Online Platforms and Running Clubs**: Social media platforms, running clubs, or online marathon communities (such as *Reddit's Running Community*, *Runkeeper*, or *Strava*) could also be used to recruit participants. These platforms allow researchers to directly contact a wide range of runners and invite them to participate.
- **Incentives**: To increase participation, participants could be incentivized with benefits like personalized marathon performance analysis, discounted race entries, or access to exclusive content (e.g., training guides).
- **Direct Invitations to Past Participants**: If accessible, invitations could also be sent to runners who have participated in recent marathons, using event data or online race databases.

What sampling approach is taken, and what are some of the trade-offs of this?

Sampling Approach: The study could use a **stratified random sampling** approach, where participants are grouped into strata based on factors such as age, gender, marathon experience, and race times. From each group, participants would be randomly selected, ensuring that all key demographics are represented in the final sample. This approach ensures diversity and gives understanding into the performance of various runner types (elite vs. recreational, young vs. older, etc.). **Trade-offs**:

- **Representativeness**: Stratified sampling ensures that the sample reflects the diversity of the marathon population, allowing for more accurate analysis across various demographic groups.

- **Higher Precision**: By ensuring that each subgroup is adequately represented, stratified sampling minimizes the potential bias that could arise from over- or under-representing certain groups.
- **Complexity**: Stratified sampling can be logistically complex, requiring careful categorization and segmentation of the sample. This can be time-consuming and might involve additional administrative effort to ensure proper stratification.
- **Cost**: If the sample size is large and the process involves targeted outreach, the cost of recruitment (e.g., incentives, outreach campaigns) could be higher than other simpler approaches, such as convenience sampling.

How is non-response handled?

Non-response occurs when individuals invited to participate in the study choose not to provide their data or drop out of the study. Addressing non-response is important for maintaining the quality and representativeness of the study.

- **Follow-up Invitations**: If a participant does not respond initially, follow-up emails, phone calls, or social media messages could be sent to encourage them to complete the survey or data collection process. These reminders can highlight the benefits of participation, such as personalized analysis of race performance.
- **Incentives**: Offering additional incentives (e.g., free access to marathon training programs or race entry discounts) can motivate reluctant participants to engage with the study.
- **Weighting for Non-Response**: In case of high non-response rates, the study could adjust the results using **statistical weighting**. If certain demographic groups (e.g., older or less experienced runners) are underrepresented due to non-response, their data could be weighted more heavily to correct for this discrepancy.
- **Post-Study Adjustment**: If certain groups are systematically underrepresented (e.g., female runners, older participants), the data analysis could use **post-hoc adjustments** to account for these disparities and ensure that the final model is as unbiased as possible.

What is good and bad about the questionnaire?

- **More Context**: Participants could be asked to report on key factors that influence their performance, such as training regimen, injury history, nutrition, sleep quality, and mental state. This contextual data could significantly improve the predictive accuracy of the model by accounting for variables not captured in race results alone.
- **Flexibility**: The questionnaire could be designed to ask open-ended or scaled questions, allowing for a broad spectrum of data to be collected. This could help uncover information that is specific to individual runners or subgroups.

- **Self-Reporting Bias**: One of the major issues with questionnaires is the potential for self-reporting bias. Participants may overestimate their training efforts, underestimate injuries, or report idealized versions of their habits, leading to inaccurate data. This can reduce the reliability of the findings and introduce errors into the predictive model.
- **Incomplete or Inconsistent Responses**: Participants may skip questions or provide inconsistent responses, making it difficult to create a clean, reliable dataset. Data cleaning processes would be necessary to handle missing or inconsistent data.
- **Recall Bias**: Runners may struggle to remember specific details about their training or race-day experiences, leading to recall bias. For example, they may not accurately report the total hours of training or the specific times they ran certain distances.

6.3 Idealized Survey

A sample survey can be found at: https://forms.gle/iRL8zHXrGtsoF3dJ8

The survey questions are detailed below as well for convenience.

Marathon Performance and Training Survey

Introduction

Thank you for participating in this survey! We are collecting data on marathon runners to better understand the factors that contribute to marathon performance. This survey should take approximately 10 minutes to complete. All responses will be kept confidential and used solely for research purposes.

Contact Information: If you have any questions about the survey or the data collection process, please contact

Survey Coordinator: Sophia Brothers
Email: sophia.brothers@mail.utoronto.ca

Section 1: Demographic Information

1. Age Group: (Select one)

- Under 20
- 20-29
- 30-39
- 40-49
- 50-59
- 60 or older

2. Gender: (Select one)

- Male
- Female
- Non-binary
- Prefer not to say
- Other: ____________

3. Biological Sex: (Select one)

- Man
- Woman
- Intersex
- Prefer not to say

3. What is your highest level of education? (Select one)

- High School or equivalent
- Some College
- Associate's Degree
- Bachelor's Degree
- Graduate or Professional Degree

4. What is your primary occupation? (Select one)

- Full-time employed
- Part-time employed
- Self-employed
- Student
- Retired
- Other (please specify): ______________________

Section 2: Marathon Experience

5. How many marathons have you completed? (Select one)

- This is my first marathon

- 1-3 marathons
- 4-6 marathons
- 7 or more marathons

6. What is your best marathon finish time? (in hours:minutes:seconds)

- Open-ended response: ______________________

7. How do you usually track your marathon performance? (Select one)

- I use a fitness tracker (e.g., Garmin, Fitbit, etc.)
- I track manually (e.g., running logs, spreadsheets)
- I use a mobile app (e.g., Strava, Runkeeper, Nike Run Club)
- I rely on race timing chips only
- Other (please specify): ____________________

Section 3: Training Habits

8. On average, how many hours per week do you spend training for a marathon? (Select one)

- 0-5 hours
- 6-10 hours
- 11-15 hours
- 16+ hours

9. What types of training do you incorporate into your marathon preparation? (Select all that apply)

- Long runs (over 15 miles)
- Tempo runs (moderate pace for endurance)
- Interval training (speed work)
- Cross-training (e.g., cycling, swimming)
- Strength training (e.g., weightlifting)
- Yoga or flexibility training
- Rest days
- Other (please specify): ______________________

10. What is your primary goal for marathon training? (Select one)

- Improve race time (speed)
- Complete the marathon (finish without injury)
- Train for an ultra marathon or longer distances
- Participate for fun/charity
- Other (please specify): ________________

11. Have you ever experienced any of the following training-related injuries? (select all that apply)

- Runner's knee
- IT band syndrome
- Shin splints
- Achilles tendonitis
- Stress fractures
- Ankle sprains
- None
- Other (please specify): ______________________

Section 4: Race-Day Factors

12. What time of day do you typically prefer to run marathons? (Select one)

- Morning (before 10 AM)
- Midday (10 AM - 3 PM)
- Afternoon (3 PM - 6 PM)
- Evening (after 6 PM)

13. Do you follow a specific nutrition plan during your marathon training or on race day? (Select one)

- Yes, I follow a strict nutrition plan
- I have a general plan but am flexible
- No, I don't follow a specific nutrition plan
- I follow a hydration-only plan

Final Section

Thank you for completing this survey! Your responses will help improve training recommendations for runners.

6.4 Additional Tables & Figures

Table 5: Summary of the marathon finishing time model, which includes key variables such as gender and country. The table presents the model coefficients along with their standard errors.

Term	Estimate	Std..Error	t.value	Pr...t..	Term_Type
Intercept	13898.563	1694.547	8.202	0.000	Linear
genderFemale	1963.963	29.654	66.229	0.000	Linear
genderOther	1601.891	311.826	5.137	0.000	Linear
iaaf_categoryAHO	2450.067	3788.981	0.647	0.518	Linear
iaaf_categoryALB	1941.249	2273.331	0.854	0.393	Linear
iaaf_categoryALG	-706.041	2036.636	-0.347	0.729	Linear
iaaf_categoryAND	-802.506	2934.878	-0.273	0.785	Linear
iaaf_categoryANG	321.291	2934.903	0.109	0.913	Linear
iaaf_categoryANT	1283.032	3788.714	0.339	0.735	Linear
iaaf_categoryARG	600.201	1708.373	0.351	0.725	Linear
iaaf_categoryARM	362.567	2588.220	0.140	0.889	Linear
iaaf_categoryARU	693.591	3788.807	0.183	0.855	Linear
iaaf_categoryAUS	1349.380	1698.106	0.795	0.427	Linear
iaaf_categoryAUT	395.074	1717.676	0.230	0.818	Linear
iaaf_categoryAZE	4729.535	2396.219	1.974	0.048	Linear
iaaf_categoryBAH	3060.501	2934.967	1.043	0.297	Linear
iaaf_categoryBAN	3240.593	2588.351	1.252	0.211	Linear
iaaf_categoryBAR	1075.716	2273.400	0.473	0.636	Linear
iaaf_categoryBEL	1251.715	1706.366	0.734	0.463	Linear
iaaf_categoryBER	-1474.061	2273.312	-0.648	0.517	Linear
iaaf_categoryBHU	-3127.660	2588.432	-1.208	0.227	Linear
iaaf_categoryBIH	779.428	2273.331	0.343	0.732	Linear
iaaf_categoryBLR	-952.755	1956.549	-0.487	0.626	Linear
iaaf_categoryBOL	3436.490	1956.636	1.756	0.079	Linear
iaaf_categoryBOT	1956.031	2935.011	0.666	0.505	Linear
iaaf_categoryBRA	534.300	1698.581	0.315	0.753	Linear
iaaf_categoryBRN	-3304.880	2588.227	-1.277	0.202	Linear
iaaf_categoryBRU	991.073	3788.835	0.262	0.794	Linear
iaaf_categoryBUL	1470.835	1907.279	0.771	0.441	Linear
iaaf_categoryCAM	-1792.784	2935.055	-0.611	0.541	Linear

Table 5: Summary of the marathon finishing time model, which includes key variables such as gender and country. The table presents the model coefficients along with their standard errors.

Term	Estimate	Std..Error	t.value	Pr...t..	Term_Type
iaaf_categoryCAN	440.070	1697.592	0.259	0.795	Linear
iaaf_categoryCAY	595.844	2273.375	0.262	0.793	Linear
iaaf_categoryCHI	320.016	1716.650	0.186	0.852	Linear
iaaf_categoryCHN	416.157	1701.324	0.245	0.807	Linear
iaaf_categoryCMR	5213.214	2934.947	1.776	0.076	Linear
iaaf_categoryCOL	555.996	1705.889	0.326	0.744	Linear
iaaf_categoryCRC	330.743	1720.010	0.192	0.848	Linear
iaaf_categoryCRO	978.011	1811.515	0.540	0.589	Linear
iaaf_categoryCUB	3004.619	2273.406	1.322	0.186	Linear
iaaf_categoryCUR	2476.043	3788.867	0.654	0.513	Linear
iaaf_categoryCYP	569.473	1956.652	0.291	0.771	Linear
iaaf_categoryCZE	304.264	1803.934	0.169	0.866	Linear
iaaf_categoryDEN	191.631	1724.691	0.111	0.912	Linear
iaaf_categoryDMA	1519.559	2934.832	0.518	0.605	Linear
iaaf_categoryDOM	1852.557	1720.761	1.077	0.282	Linear
iaaf_categoryECU	222.955	1723.771	0.129	0.897	Linear
iaaf_categoryEGY	2522.537	1921.470	1.313	0.189	Linear
iaaf_categoryESA	1550.410	1824.983	0.850	0.396	Linear
iaaf_categoryESP	79.320	1699.283	0.047	0.963	Linear
iaaf_categoryEST	419.637	1811.553	0.232	0.817	Linear
iaaf_categoryETH	-1911.091	1921.388	-0.995	0.320	Linear
iaaf_categoryFIJ	2829.413	3788.772	0.747	0.455	Linear
iaaf_categoryFIN	-12.371	1768.204	-0.007	0.994	Linear
iaaf_categoryFRA	964.275	1695.981	0.569	0.570	Linear
iaaf_categoryGAM	3425.591	3788.807	0.904	0.366	Linear
iaaf_categoryGBR	733.388	1696.359	0.432	0.666	Linear
iaaf_categoryGEO	-451.588	3788.685	-0.119	0.905	Linear
iaaf_categoryGER	959.991	1697.010	0.566	0.572	Linear
iaaf_categoryGHA	-1660.468	3788.856	-0.438	0.661	Linear
iaaf_categoryGRE	146.601	1791.496	0.082	0.935	Linear
iaaf_categoryGRN	1043.355	3789.274	0.275	0.783	Linear
iaaf_categoryGUA	1032.157	1729.803	0.597	0.551	Linear
iaaf_categoryGUD	3784.233	3788.895	0.999	0.318	Linear
iaaf_categoryGUM	3659.554	3788.857	0.966	0.334	Linear
iaaf_categoryGUY	2277.875	2273.889	1.002	0.316	Linear
iaaf_categoryHAI	9336.025	2273.570	4.106	0.000	Linear
iaaf_categoryHKG	667.923	1718.198	0.389	0.697	Linear

Table 5: Summary of the marathon finishing time model, which includes key variables such as gender and country. The table presents the model coefficients along with their standard errors.

Term	Estimate	Std..Error	t.value	Pr...t..	Term_Type
iaaf_categoryHON	1409.680	1830.208	0.770	0.441	Linear
iaaf_categoryHUN	1192.440	1781.436	0.669	0.503	Linear
iaaf_categoryINA	3363.264	1713.251	1.963	0.050	Linear
iaaf_categoryIND	3135.782	1703.501	1.841	0.066	Linear
iaaf_categoryIRI	3941.520	2075.284	1.899	0.058	Linear
iaaf_categoryIRL	651.394	1701.800	0.383	0.702	Linear
iaaf_categoryIRQ	627.790	3788.667	0.166	0.868	Linear
iaaf_categoryISL	-204.266	1791.392	-0.114	0.909	Linear
iaaf_categoryISR	677.109	1705.996	0.397	0.691	Linear
iaaf_categoryITA	1706.363	1695.804	1.006	0.314	Linear
iaaf_categoryJAM	3995.560	1873.418	2.133	0.033	Linear
iaaf_categoryJER	1228.400	2273.415	0.540	0.589	Linear
iaaf_categoryJPN	836.911	1703.927	0.491	0.623	Linear
iaaf_categoryKAZ	-383.024	1830.223	-0.209	0.834	Linear
iaaf_categoryKEN	-2042.009	1807.532	-1.130	0.259	Linear
iaaf_categoryKGZ	276.234	2934.969	0.094	0.925	Linear
iaaf_categoryKOR	1216.739	1716.324	0.709	0.478	Linear
iaaf_categoryKOS	-768.640	2934.821	-0.262	0.793	Linear
iaaf_categoryKSA	3546.430	2075.209	1.709	0.087	Linear
iaaf_categoryKUW	1693.176	2396.423	0.707	0.480	Linear
iaaf_categoryLAT	-318.619	1937.741	-0.164	0.869	Linear
iaaf_categoryLCA	6950.961	2588.301	2.686	0.007	Linear
iaaf_categoryLIB	1754.339	1978.870	0.887	0.375	Linear
iaaf_categoryLIE	-837.675	2187.446	-0.383	0.702	Linear
iaaf_categoryLTU	-760.230	1811.493	-0.420	0.675	Linear
iaaf_categoryLUX	1963.530	1883.316	1.043	0.297	Linear
iaaf_categoryMAC	-790.235	2588.235	-0.305	0.760	Linear
iaaf_categoryMAR	350.908	1737.977	0.202	0.840	Linear
iaaf_categoryMAS	2317.788	1754.979	1.321	0.187	Linear
iaaf_categoryMAW	6538.128	3788.849	1.726	0.084	Linear
iaaf_categoryMDA	-231.520	2273.266	-0.102	0.919	Linear
iaaf_categoryMEX	954.235	1697.040	0.562	0.574	Linear
iaaf_categoryMGL	574.618	1937.628	0.297	0.767	Linear
iaaf_categoryMKD	1417.591	2934.768	0.483	0.629	Linear
iaaf_categoryMLT	835.559	2187.610	0.382	0.702	Linear
iaaf_categoryMNE	689.194	2396.381	0.288	0.774	Linear
iaaf_categoryMON	-151.943	2934.825	-0.052	0.959	Linear

Table 5: Summary of the marathon finishing time model, which includes key variables such as gender and country. The table presents the model coefficients along with their standard errors.

Term	Estimate	Std..Error	t.value	Pr...t..	Term_Type
iaaf_categoryMRI	10446.763	3788.704	2.757	0.006	Linear
iaaf_categoryMTN	4548.926	3788.786	1.201	0.230	Linear
iaaf_categoryMYA	1835.148	3788.946	0.484	0.628	Linear
iaaf_categoryNA	2912.359	3792.310	0.768	0.443	Linear
iaaf_categoryNAM	1675.073	3788.835	0.442	0.658	Linear
iaaf_categoryNCA	1009.368	1937.711	0.521	0.602	Linear
iaaf_categoryNED	884.842	1697.979	0.521	0.602	Linear
iaaf_categoryNEP	2846.108	1921.391	1.481	0.139	Linear
iaaf_categoryNGR	5662.102	2036.426	2.780	0.005	Linear
iaaf_categoryNIG	3885.350	3788.880	1.025	0.305	Linear
iaaf_categoryNOR	315.084	1705.331	0.185	0.853	Linear
iaaf_categoryNZL	2298.011	1715.496	1.340	0.180	Linear
iaaf_categoryPAK	2940.249	1800.604	1.633	0.102	Linear
iaaf_categoryPAN	673.099	1786.134	0.377	0.706	Linear
iaaf_categoryPAR	2278.280	1807.519	1.260	0.208	Linear
iaaf_categoryPER	885.714	1722.014	0.514	0.607	Linear
iaaf_categoryPHI	3853.325	1707.940	2.256	0.024	Linear
iaaf_categoryPLE	670.739	2588.220	0.259	0.796	Linear
iaaf_categoryPOL	373.694	1704.379	0.219	0.826	Linear
iaaf_categoryPOR	-310.287	1726.254	-0.180	0.857	Linear
iaaf_categoryPUR	2932.722	1750.066	1.676	0.094	Linear
iaaf_categoryPYF	250.805	3788.817	0.066	0.947	Linear
iaaf_categoryQAT	-223.732	3788.887	-0.059	0.953	Linear
iaaf_categoryREU	1143.871	2273.631	0.503	0.615	Linear
iaaf_categoryROM	715.808	1815.692	0.394	0.693	Linear
iaaf_categoryRSA	1555.399	1714.294	0.907	0.364	Linear
iaaf_categoryRUS	293.778	1752.943	0.168	0.867	Linear
iaaf_categorySIN	1894.121	1739.656	1.089	0.276	Linear
iaaf_categorySLO	-635.859	1842.118	-0.345	0.730	Linear
iaaf_categorySMR	2305.150	2588.318	0.891	0.373	Linear
iaaf_categorySRB	-18.026	1856.429	-0.010	0.992	Linear
iaaf_categorySRI	3917.483	2588.586	1.513	0.130	Linear
iaaf_categorySUI	697.701	1702.596	0.410	0.682	Linear
iaaf_categorySVK	369.556	1800.474	0.205	0.837	Linear
iaaf_categorySWE	1214.933	1704.604	0.713	0.476	Linear
iaaf_categorySWZ	7016.448	3789.186	1.852	0.064	Linear
iaaf_categorySYR	2494.682	2396.334	1.041	0.298	Linear

Table 5: Summary of the marathon finishing time model, which includes key variables such as gender and country. The table presents the model coefficients along with their standard errors.

Term	Estimate	Std..Error	t.value	Pr...t..	Term_Type
iaaf_categoryTAN	2937.820	2396.462	1.226	0.220	Linear
iaaf_categoryTHA	2596.083	1710.757	1.518	0.129	Linear
iaaf_categoryTJK	-958.606	3788.733	-0.253	0.800	Linear
iaaf_categoryTKM	7964.080	3789.024	2.102	0.036	Linear
iaaf_categoryTKS	3889.028	2934.958	1.325	0.185	Linear
iaaf_categoryTOG	4249.812	3788.874	1.122	0.262	Linear
iaaf_categoryTPE	1252.888	1714.444	0.731	0.465	Linear
iaaf_categoryTRI	3128.996	1856.218	1.686	0.092	Linear
iaaf_categoryTUN	1843.912	2396.755	0.769	0.442	Linear
iaaf_categoryTUR	1463.802	1786.158	0.820	0.412	Linear
iaaf_categoryUAE	4949.387	2588.324	1.912	0.056	Linear
iaaf_categoryUGA	755.734	2036.547	0.371	0.711	Linear
iaaf_categoryUKR	369.674	1733.937	0.213	0.831	Linear
iaaf_categoryURU	771.917	1755.029	0.440	0.660	Linear
iaaf_categoryUSA	1859.961	1694.525	1.098	0.272	Linear
iaaf_categoryUZB	1475.017	2396.292	0.616	0.538	Linear
iaaf_categoryVEN	1577.256	1719.374	0.917	0.359	Linear
iaaf_categoryVIE	2120.320	2004.835	1.058	0.290	Linear
iaaf_categoryVIN	11446.494	3788.816	3.021	0.003	Linear
iaaf_categoryZAM	4382.916	2934.904	1.493	0.135	Linear
iaaf_categoryZIM	1023.408	1978.673	0.517	0.605	Linear

Table 6: Summary of the marathon finishing time model, which includes smooth variables such as age and race count The table presents the model coefficients along with their standard errors.

Term	edf	Ref.df	F	p.value	Term_Type
Smooth Term - Age	8.633	8.919	541.776	0	Smooth
Smooth Term - Race Count	8.385	8.855	8.667	0	Smooth

References

Arel-Bundock, Vincent. 2023. *Modelsummary: Beautiful and Customizable Model Summaries and Tables in r*. https://CRAN.R-project.org/package=modelsummary.

Cheuvront, Samuel N, Robert Carter, Keith C Deruisseau, and Robert J Moffatt. 2005. "Running Performance Differences Between Men and Women: An Update." *Sports Medicine* 35 (12): 1017–24. https://doi.org/10.2165/00007256-200535120-00002.

Connick, Mark J, Emma M Beckman, and Sean M Tweedy. 2015. "Relative Age Affects Marathon Performance in Male and Female Athletes." *Journal of Sports Science & Medicine* 14 (3): 669–74. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4541133/.

Hillier, Bianca. 2024. "More People Are Running Marathons Than Ever Before. Why?" *The World*. https://theworld.org/stories/2024/05/23/more-people-are-running-marathons-than-ever-before-why.

Hovde, Joe. 2024. "NYC Marathon Finishers Dataset, 2024." https://www.data-is-plural.com/archive/2024-11-13-edition/.

Kuhn, Max. 2023. *Caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret.

Müller, Kirill. 2023. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, et al. 2023. *Arrow: Integration to 'Apache Arrow'*. https://CRAN.R-project.org/package=arrow.

Spinu, Vitalie, Garrett Grolemund, and Hadley Wickham. 2023. *Lubridate: Make Dealing with Dates a Little Easier*. https://CRAN.R-project.org/package=lubridate.

Wickham, Hadley. 2023a. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

———. 2023b. *Testthat: Unit Testing for r*. https://cran.r-project.org/web/packages/testthat/index.html.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, Dewey Dunnington, and Teun van den Brand. 2023. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, Romain Francois, and Jennifer Bryan. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Wilber, Randall L., and Yannis P. Pitsiladis. 2012. "Kenyan and Ethiopian Distance Runners:

What Makes Them so Good?" *International Journal of Sports Physiology and Performance* 7 (2): 92–102. https://doi.org/10.1123/ijspp.7.2.92.

Zhu, Hao. 2023. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.