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

Age, gender, nationality, and race experience are the strongest factors influencing performance.

Sophia Brothers

November 26, 2024

This paper develops a linear 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 insights into the characteristics that influence marathon performance, contributing to a deeper understanding of endurance running. This research provides valuable insights 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 in recent years, attracting individuals from diverse backgrounds to participate in races of varying distances and difficulty levels (Hillier 2024). Despite the increasing number of participants, predicting marathon performance remains a challenging task due to the multitude of factors influencing finishing times. These factors include age, gender, experience, and training, all of which contribute to the overall race outcome. However, while existing studies have explored marathon times, there remains a gap in understanding the quantitative relationships between these variables and how they influence performance.

This paper aims to fill this gap by developing a linear regression model to predict marathon finishing times based on variables such as age, gender, country, and previous race experience. The primary estimand of this study is the prediction of a runner's overall marathon finishing time, in seconds, as a function of their age, gender, country, and race count. By estimating this relationship, the model quantifies the effect of these variables on marathon performance and identifies the key factors that influence race outcomes.

The analysis revealed that age and gender are significant predictors of marathon finishing times, with older runners generally taking longer to finish, and male runners achieving faster times. These findings contribute to the broader understanding of marathon performance and can assist in setting realistic expectations for future participants based on their individual characteristics.

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 linear regression model, 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 comprehensive profile of marathon participants, enabling the exploration of how different factors contribute to marathon outcomes.

We used R (R Core Team 2023) for data cleaning and analysis, leveraging 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.

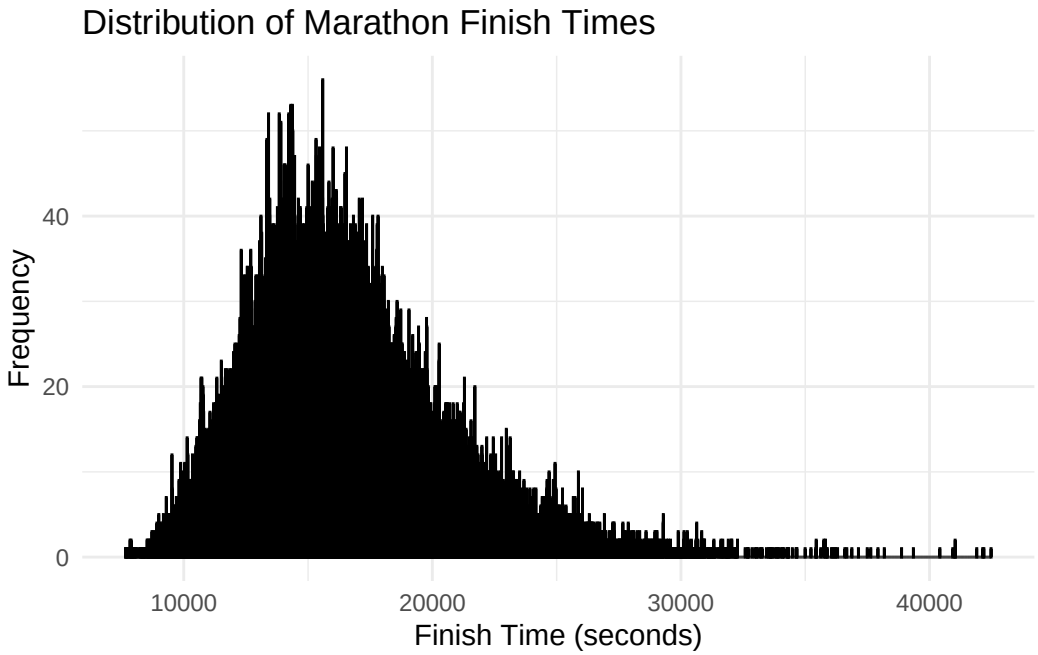

Figure 1: Distribution of marathon finish times, highlighting the concentration around the 4-hour mark. The skew towards longer finish times reflects the diverse abilities of participants.

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.

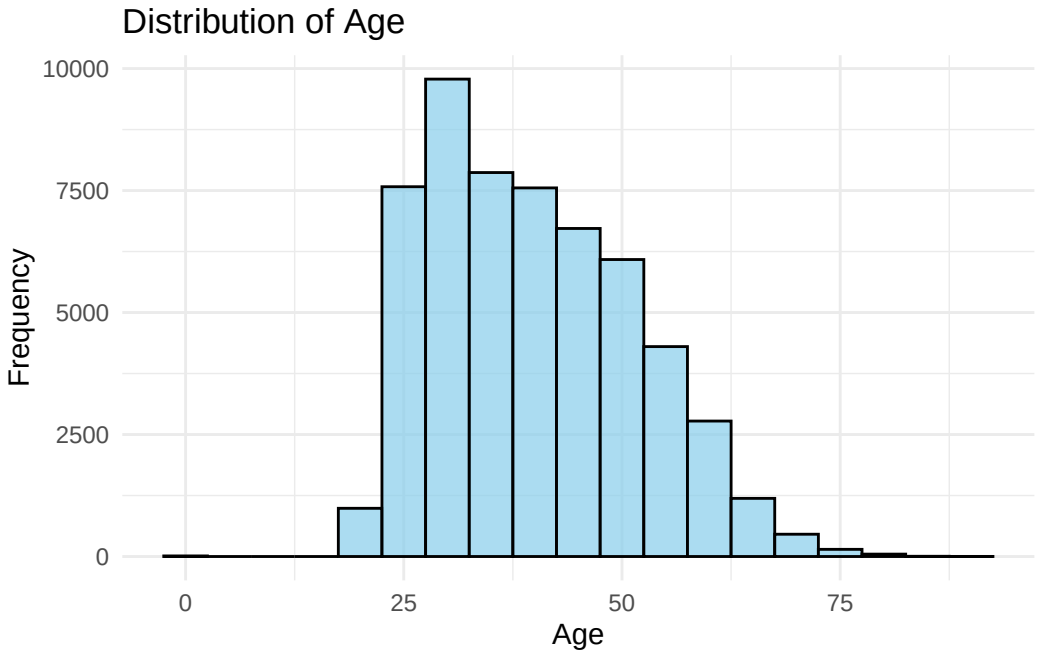

Figure 2: Distribution of runners' ages, showing a wide range of participation with a peak in the 30-40 age group, a common range for competitive runners.

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

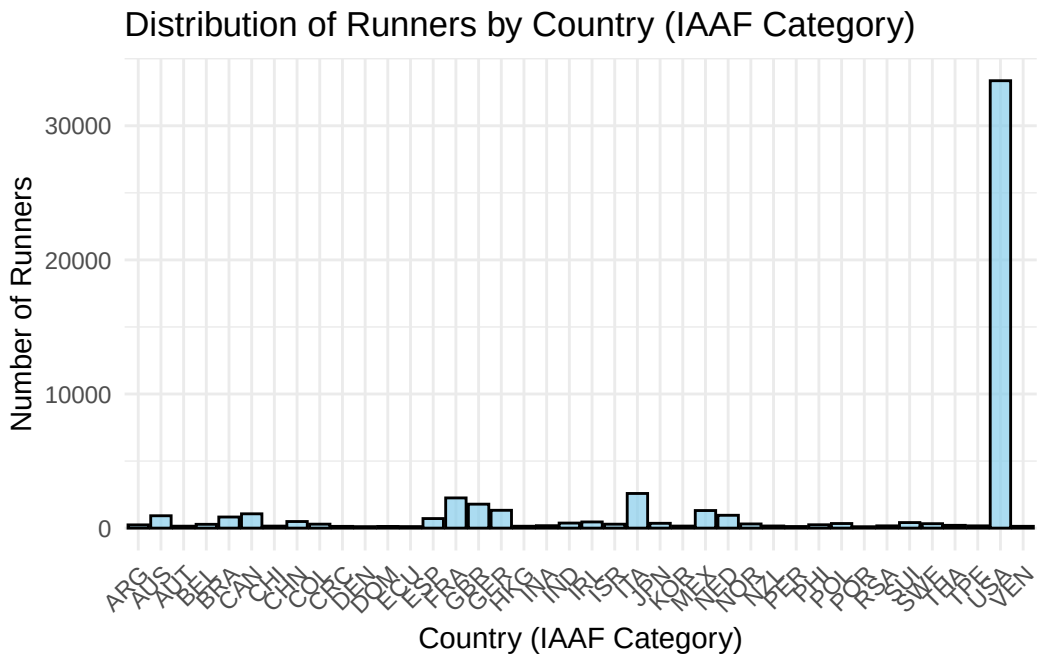

Figure 3: Distribution of marathon participants by country, limited to those with more than 100 runners. The United States has the highest number of participants, followed by several other countries with a smaller but significant representation.

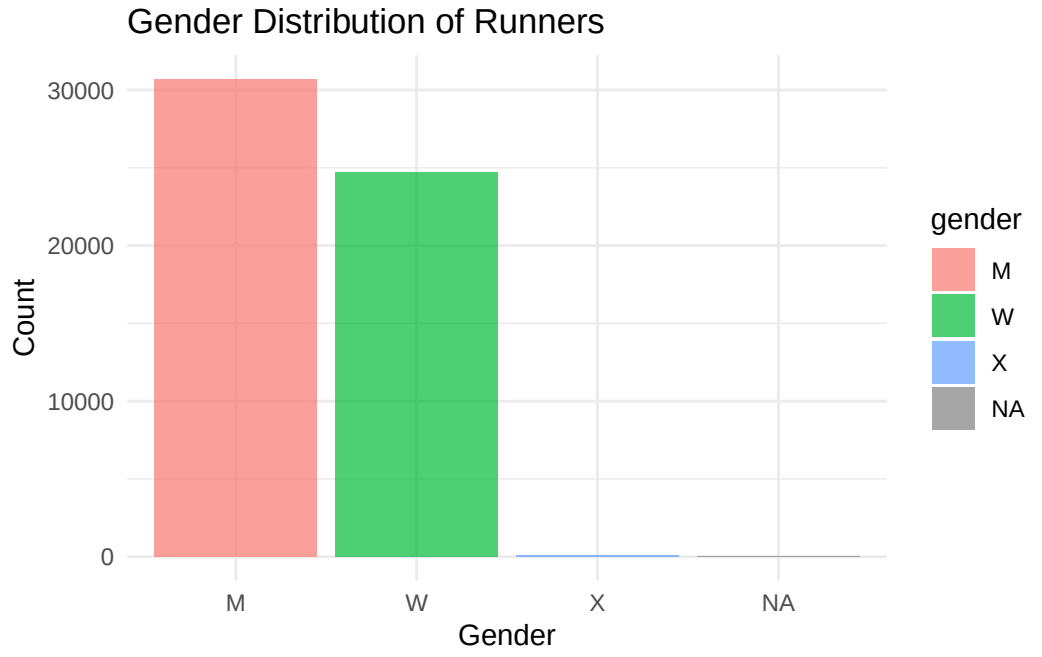

Figure 4: Gender distribution among marathon participants, revealing a relatively even split, with a slight predominance of male runners.

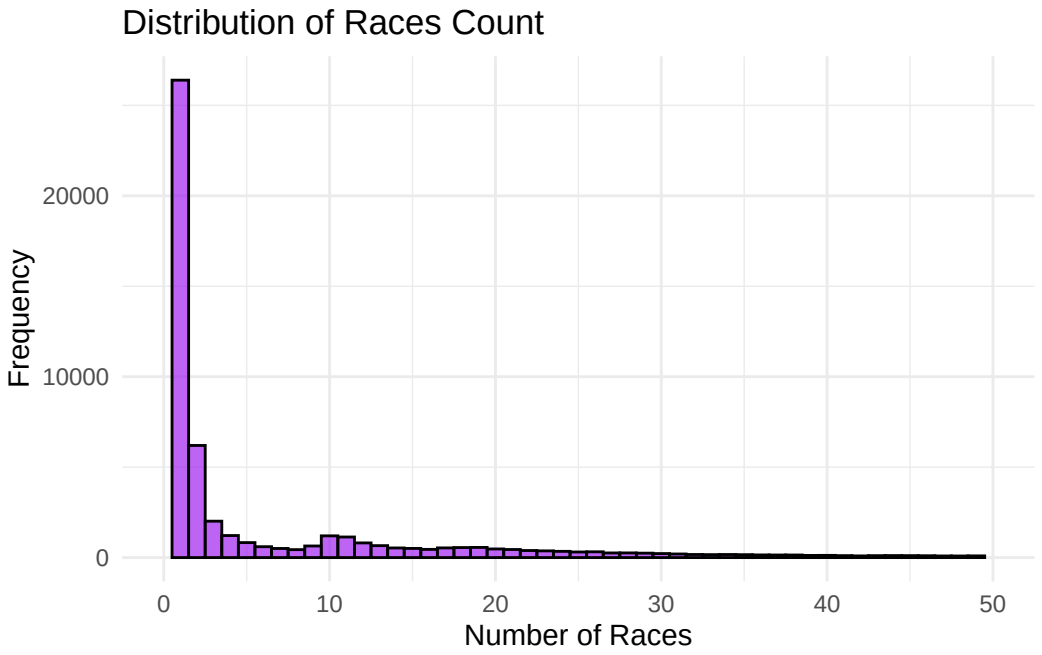

Figure 5: Distribution of the number of races completed by runners, indicating that most participants are relatively inexperienced, with a smaller number of highly experienced runners.

contribute to overall race performance, helping to provide insights 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 linear model 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 linear regression model is specified as follows:

$$overall\_time\_seconds_i = \beta_0 + \beta_1 \cdot age_i + \beta_2 \cdot gender_i + \beta_3 \cdot race\_count_i + \beta_4 \cdot iaaf\_category + \epsilon_i$$

Where:

- `overall_time_seconds` is the marathon finishing time in seconds for the i-th runner.
- `age` is the runner's age at the time of the race
- `gender` is the runner's gender (treated as a categorical variable).
- `race_count` is the total number of marathons the runner has participated in prior to the race.
- `iaaf_category` is the country the runner is from under IAAF standards.
- epsilon is the error term, assumed to be normally distributed with a mean of 0 and constant variance $\epsilon_i \sim N(0, \sigma^2)$

The model assumes that the predictors have a linear relationship with the response variable, i.e., the overall time is linearly related to the selected features. We are interested in estimating the coefficients β_1, β_2,...,β_4 which represent the effect of each predictor on the outcome.

3.2 Model Justification

The linear regression model was chosen for several reasons. Firstly, a linear regression model is straightforward and easy to interpret. Each coefficient directly indicates how much the marathon finishing time changes with a one-unit increase in the corresponding predictor, assuming all other variables remain constant.

The model accommodates both continuous predictors (like age and race count) and categorical variables (such as gender and IAAF country classification). This flexibility makes it well-suited for the dataset.

Additionally, previous studies in sports science consistently show that factors like age, gender, nationality, and running experience significantly affect marathon performance Wilber and Pitsiladis (2012). Including these predictors aligns with established domain knowledge, ensuring the model is based on sound reasoning.

Models like random forests and generalized additive models (GAMs) were also explored. These can capture more complex relationships but are harder to interpret and communicate. Since interpretability is a priority for this analysis, linear regression was preferred. Alternative models may be more appropriate if predictive power is the main goal, but simplicity and transparency make linear regression the better choice for the current objectives.

3.3 Assumptions and Limitations

The linear regression model relies on several key assumptions. First, it assumes a linear relationship between the predictors and the outcome variable, meaning that a unit change in a predictor corresponds to a consistent change in the response variable. Second, the residuals of the model must be independent to ensure that the statistical inferences drawn, such as significance tests and confidence intervals, are valid. Third, the variance of the residuals (errors) should remain constant across all levels of the predictors, an assumption known as homoscedasticity. Fourth, the errors should follow a normal distribution, which is critical for accurate hypothesis testing and confidence interval estimation. Lastly, the model assumes no perfect multicollinearity among the predictors, as highly correlated variables can obscure the individual effects of each predictor on the outcome.

Regardless of these assumptions, the model has limitations. It may not effectively capture complex or non-linear relationships between predictors and the outcome, which could lead to an oversimplified interpretation of the data. For example, the effect of age on marathon times may not be linear across all age groups, and a non-linear model might better capture this relationship. Interaction effects, where certain combinations of predictors interact to influence finishing times, are not accounted for in this model. For instance, the combined effect of experience (race count) and age might have a more significant impact on marathon performance than either variable alone. Additionally, outliers—such as unusually high or low marathon times—can disproportionately affect the results, potentially skewing the analysis.

Other external factors that may influence marathon performance, but are not included in this model, could limit its explanatory power. Environmental factors like weather conditions on race day, course difficulty (e.g., elevation changes), and runner health or injuries are not considered in this analysis, though they could significantly affect finishing times. Moreover, the model assumes that all runners have similar training levels, but differences in training intensity, diet, and sleep patterns could also impact performance. The model does not address these outstanding factors, which may limit its applicability in specific scenarios. These assumptions and limitations should be carefully considered when interpreting the model's results and applying them to broader contexts.

3.4 Model Validation

Model validation is an essential part of ensuring that the model generalizes well to new, unseen data. To validate the performance of the linear 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 3416.17. The RMSE provides an estimate of the average error between the predicted and actual values of the outcome variable, meaning that the average error between the model and actual time is $\pm$ 3416.17 seconds. A lower RMSE indicates better predictive performance.

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.

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

Average Finish Time (minutes)	Average Age (years)	Average Pace (minutes/mile)	Total Participants
271.8	39.89	10.37	55524

Figure 6 compares the predicted and actual marathon times across different age groups. The plot reveals 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	10 766.379
	(1711.551)
GenderFemale	1957.609
	(29.874)
GenderOther	1655.546
	(314.792)
Age	82.683
	(1.358)
Race Count	4.495
	(0.513)
Num.Obs.	55 512
R2	0.152
R2 Adj.	0.149
AIC	1 061 188.0
BIC	1 062 660.5
Log.Lik.	−530 429.004
F	60.851
RMSE	3416.17

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.

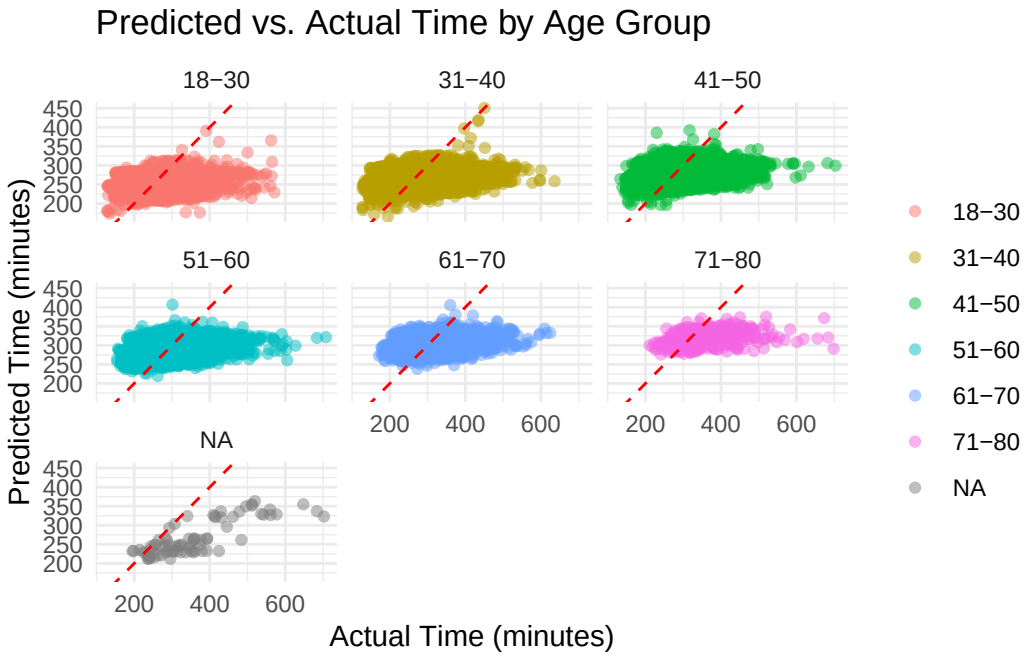

Figure 6: Predicted vs Actual Marathon Finishing Times by Age Group. The scatter plot shows how well the predicted finishing times compare with the actual times across different age groups, with a reference line indicating perfect predictions.

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**

    - **Actual Best Group**: Age 28 had the fastest mean actual marathon time of 15,570.83 seconds, with a participant count of 2,042.

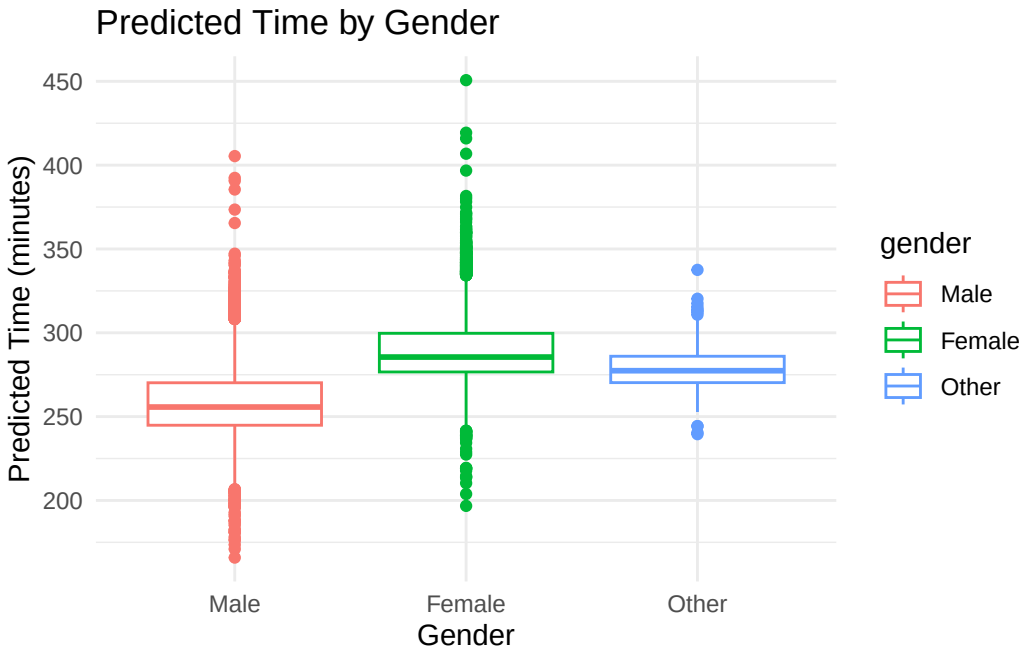

Figure 7: Predicted Marathon Finishing Times by Gender. The boxplot displays the distribution of predicted marathon times for different genders, showing median values and variability in finishing times.

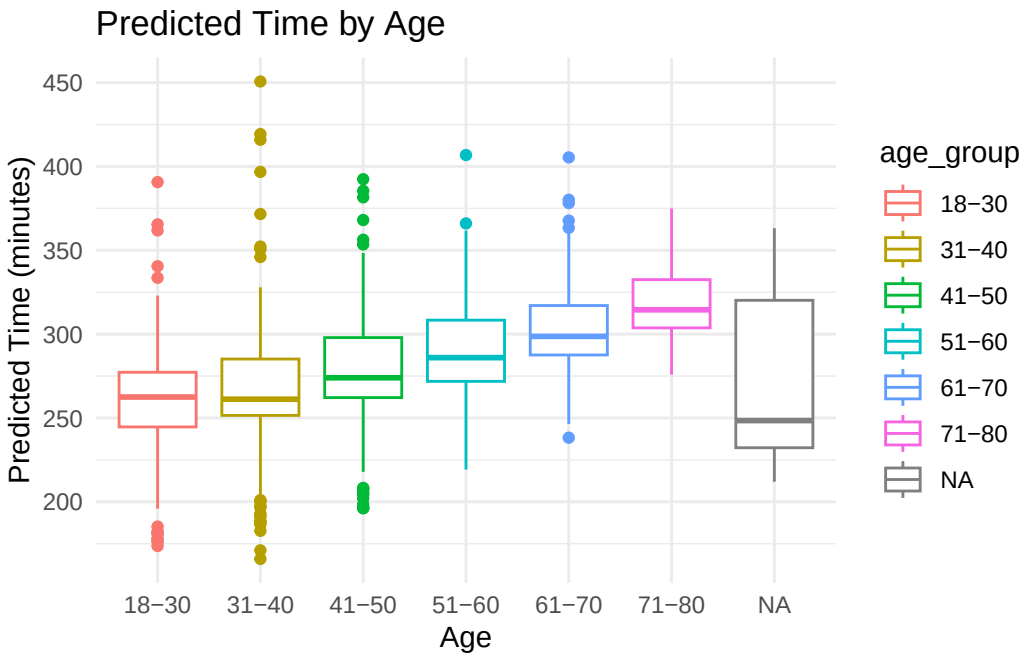

Figure 8: Predicted Marathon Finishing Times by Age Group. This boxplot visualizes the predicted marathon times across various age groups, showing how age influences the predicted finishing times.

- **Predicted Best Group**: The predictive model identified Age 18 as the best-performing group, with a faster predicted mean time of 14,340.60 seconds.

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 10 races as the best-performing group, with a slightly slower mean time of 16,042.84 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 12,079.54 seconds (predicted).

Table 3: Actual Marathon Finishing Times by Group. 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: Predicted Marathon Finishing Times by Group. 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	18	14400.66
Gender	Male	15486.41
Races Count	10	16049.96
IAAF Category	KEN	12157.34

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 valuable insights 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. Moreover, the model was able to predict marathon finishing times with a high degree of accuracy, as evidenced by the close alignment of actual and predicted times in the scatter plot.

The analysis also provided insight into the variability in predicted times, especially through the inclusion of prediction intervals, which revealed the range of uncertainty surrounding each prediction. This aspect of the analysis helps highlight the potential for error in any model and the need to consider prediction intervals when interpreting the results. In particular, the spread of predicted times by age and gender demonstrates the complexity of predicting marathon times for individuals based on demographic data alone.

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 revealed 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 of the Model:

While the model provided valuable insights 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. For example, if the majority of participants in the dataset come from a particular geographic region or socio-economic background, the model may not perform as well for runners from diverse backgrounds. In this case, since the data is solely from the NYC marathon, a significant proportion of the runners are from the NYC metro area and the USA in general.

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).

5.5 Next Steps:

Future research should explore 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 more comprehensive view of the factors influencing marathon times. Incorporating these variables would likely enhance the model's accuracy and offer more personalized predictions.

Additionally, current models typically assume linear relationships between predictors and outcomes, but marathon times may be influenced by more complex interactions that could be better captured by more advanced modeling techniques like random forests or gradient boosting.

Another area for future research could involve exploring 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.

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 critical 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

6.2.1 1. 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.

6.2.2 2. 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.

6.2.3 3. 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 insights 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 insights 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.

6.2.4 4. 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.

6.2.5 5. 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 insights that are 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

Certainly! Here's an example **survey questionnaire** designed to collect demographic and training data from marathon runners. The structure includes an **introductory section**, followed by **well-constructed questions** ordered logically, and ends with a **thank-you message**. The question types are varied, ensuring the survey is engaging and collects the necessary data for building a more accurate marathon performance model.

6.3.1 Marathon Performance and Training Survey

6.3.2 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

6.3.3 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

3. **Biological Sex:** (Select one)
   - Male
   - Female
   - 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): _____

6.3.4 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): _____

6.3.5 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?

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

6.3.6 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

6.3.7 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 age, gender, pace, race count, and age-graded performance. The table presents the model coefficients along with their standard errors.

	(1)
(Intercept)	10 766.379
	(1711.551)
age	82.683
	(1.358)
genderFemale	1957.609
	(29.874)
genderOther	1655.546
	(314.792)
races_count	4.495
	(0.513)
iaaf_categoryAHO	2069.575
	(3825.179)
iaaf_categoryALB	1626.978
	(2295.082)
iaaf_categoryALG	−1097.867
	(2056.040)
iaaf_categoryAND	−1110.114
	(2962.955)
iaaf_categoryANG	−215.028
	(2962.971)
iaaf_categoryANT	1003.326
	(3825.134)

iaaf_categoryARG	295.568
	(1724.699)
iaaf_categoryARM	163.625
	(2613.031)
iaaf_categoryARU	247.890
	(3825.055)
iaaf_categoryAUS	1073.104
	(1714.350)
iaaf_categoryAUT	134.646
	(1734.116)
iaaf_categoryAZE	4463.495
	(2419.187)
iaaf_categoryBAH	3336.684
	(2963.050)
iaaf_categoryBAN	2769.370
	(2613.101)
iaaf_categoryBAR	705.852
	(2295.069)
iaaf_categoryBEL	1000.945
	(1722.674)
iaaf_categoryBER	−1521.067
	(2295.163)
iaaf_categoryBHU	−3379.531
	(2613.083)
iaaf_categoryBIH	665.712
	(2295.070)
iaaf_categoryBLR	−1263.464
	(1975.261)
iaaf_categoryBOL	3158.657
	(1975.291)
iaaf_categoryBOT	1494.054
	(2962.980)
iaaf_categoryBRA	184.132
	(1714.784)
iaaf_categoryBRN	−3486.291
	(2613.012)
iaaf_categoryBRU	612.182
	(3825.112)
iaaf_categoryBUL	1158.345
	(1925.382)
iaaf_categoryCAM	−2259.417
	(2962.963)
iaaf_categoryCAN	183.532

	(1713.840)
iaaf_categoryCAY	198.382
	(2295.138)
iaaf_categoryCHI	−38.034
	(1733.046)
iaaf_categoryCHN	148.956
	(1717.574)
iaaf_categoryCMR	4850.997
	(2962.917)
iaaf_categoryCOL	219.503
	(1722.137)
iaaf_categoryCRC	−66.045
	(1736.369)
iaaf_categoryCRO	609.992
	(1828.765)
iaaf_categoryCUB	2877.696
	(2295.115)
iaaf_categoryCUR	2463.656
	(3825.096)
iaaf_categoryCYP	216.130
	(1975.378)
iaaf_categoryCZE	−69.235
	(1821.141)
iaaf_categoryDEN	113.751
	(1741.234)
iaaf_categoryDMA	1084.475
	(2962.955)
iaaf_categoryDOM	1555.684
	(1737.203)
iaaf_categoryECU	−97.970
	(1740.199)
iaaf_categoryEGY	2161.198
	(1939.766)
iaaf_categoryESA	1271.913
	(1842.416)
iaaf_categoryESP	−259.523
	(1715.506)
iaaf_categoryEST	13.550
	(1828.755)
iaaf_categoryETH	−2008.977
	(1939.796)
iaaf_categoryFIJ	2588.692
	(3825.144)

iaaf_categoryFIN	−307.857
	(1785.087)
iaaf_categoryFRA	694.805
	(1712.196)
iaaf_categoryGAM	2979.890
	(3825.055)
iaaf_categoryGBR	457.393
	(1712.575)
iaaf_categoryGEO	−623.013
	(3825.057)
iaaf_categoryGER	708.310
	(1713.254)
iaaf_categoryGHA	−2256.280
	(3825.126)
iaaf_categoryGRE	−221.098
	(1808.490)
iaaf_categoryGRN	758.770
	(3825.196)
iaaf_categoryGUA	724.999
	(1746.295)
iaaf_categoryGUD	3320.989
	(3825.145)
iaaf_categoryGUM	3207.499
	(3825.112)
iaaf_categoryGUY	2130.045
	(2295.299)
iaaf_categoryHAI	8967.952
	(2295.102)
iaaf_categoryHKG	318.316
	(1734.619)
iaaf_categoryHON	1121.194
	(1847.694)
iaaf_categoryHUN	1007.219
	(1798.418)
iaaf_categoryINA	2944.708
	(1729.556)
iaaf_categoryIND	2821.690
	(1719.731)
iaaf_categoryIRI	3640.181
	(2095.080)
iaaf_categoryIRL	348.536
	(1718.057)
iaaf_categoryIRQ	480.414

	(3825.068)
iaaf_categoryISL	−520.941
	(1808.526)
iaaf_categoryISR	290.131
	(1722.250)
iaaf_categoryITA	1472.176
	(1712.029)
iaaf_categoryJAM	3652.485
	(1891.245)
iaaf_categoryJER	706.294
	(2295.049)
iaaf_categoryJPN	646.934
	(1720.240)
iaaf_categoryKAZ	−716.216
	(1847.703)
iaaf_categoryKEN	−2392.834
	(1824.817)
iaaf_categoryKGZ	−228.916
	(2962.965)
iaaf_categoryKOR	1056.113
	(1732.758)
iaaf_categoryKOS	−1111.267
	(2962.954)
iaaf_categoryKSA	3180.758
	(2095.067)
iaaf_categoryKUW	1122.492
	(2419.228)
iaaf_categoryLAT	−781.152
	(1956.205)
iaaf_categoryLCA	6714.216
	(2613.055)
iaaf_categoryLIB	1309.602
	(1997.691)
iaaf_categoryLIE	−1113.582
	(2208.437)
iaaf_categoryLTU	−1170.934
	(1828.766)
iaaf_categoryLUX	1635.027
	(1901.339)
iaaf_categoryMAC	−978.255
	(2613.035)
iaaf_categoryMAR	−0.908
	(1754.584)

iaaf_categoryMAS	2044.010
	(1771.775)
iaaf_categoryMAW	5908.273
	(3825.117)
iaaf_categoryMDA	−421.387
	(2295.052)
iaaf_categoryMEX	661.807
	(1713.252)
iaaf_categoryMGL	223.691
	(1956.173)
iaaf_categoryMKD	1242.339
	(2962.888)
iaaf_categoryMLT	487.322
	(2208.438)
iaaf_categoryMNE	409.174
	(2419.196)
iaaf_categoryMON	−421.925
	(2963.039)
iaaf_categoryMRI	10 189.644
	(3825.129)
iaaf_categoryMTN	4301.547
	(3825.114)
iaaf_categoryMYA	1737.331
	(3825.176)
iaaf_categoryN/A	5324.836
	(3825.213)
iaaf_categoryNAM	1296.182
	(3825.112)
iaaf_categoryNCA	714.684
	(1956.180)
iaaf_categoryNED	685.963
	(1714.225)
iaaf_categoryNEP	2514.318
	(1939.707)
iaaf_categoryNGR	5468.994
	(2055.930)
iaaf_categoryNIG	3371.792
	(3825.124)
iaaf_categoryNOR	18.563
	(1721.620)
iaaf_categoryNZL	2150.035
	(1731.902)
iaaf_categoryPAK	2517.961

	(1817.733)
iaaf_categoryPAN	266.234
	(1803.166)
iaaf_categoryPAR	1941.701
	(1824.809)
iaaf_categoryPER	596.014
	(1738.457)
iaaf_categoryPHI	3526.665
	(1724.236)
iaaf_categoryPLE	369.891
	(2613.043)
iaaf_categoryPOL	17.585
	(1720.611)
iaaf_categoryPOR	−662.282
	(1742.707)
iaaf_categoryPUR	2539.963
	(1766.748)
iaaf_categoryPYF	−262.793
	(3825.057)
iaaf_categoryQAT	−176.625
	(3825.197)
iaaf_categoryREU	1373.807
	(2295.247)
iaaf_categoryROM	375.717
	(1832.988)
iaaf_categoryRSA	1241.497
	(1730.659)
iaaf_categoryRUS	−63.710
	(1769.648)
iaaf_categorySIN	1573.728
	(1756.285)
iaaf_categorySLO	−821.050
	(1859.758)
iaaf_categorySMR	1897.028
	(2613.109)
iaaf_categorySRB	−380.950
	(1874.179)
iaaf_categorySRI	3398.834
	(2613.189)
iaaf_categorySUI	483.874
	(1718.893)
iaaf_categorySVK	41.914
	(1817.677)

iaaf_categorySWE	1011.961
	(1720.930)
iaaf_categorySWZ	6486.359
	(3825.183)
iaaf_categorySYR	2053.265
	(2419.193)
iaaf_categoryTAN	2480.712
	(2419.263)
iaaf_categoryTHA	2193.987
	(1727.042)
iaaf_categoryTJK	−1124.394
	(3825.119)
iaaf_categoryTKM	7755.779
	(3825.102)
iaaf_categoryTKS	3499.456
	(2963.016)
iaaf_categoryTOG	3724.354
	(3825.135)
iaaf_categoryTPE	906.594
	(1730.818)
iaaf_categoryTRI	2898.360
	(1873.959)
iaaf_categoryTUN	2457.266
	(2419.228)
iaaf_categoryTUR	1161.787
	(1803.214)
iaaf_categoryUAE	4581.328
	(2613.048)
iaaf_categoryUGA	271.732
	(2055.965)
iaaf_categoryUKR	40.204
	(1750.459)
iaaf_categoryURU	439.441
	(1771.801)
iaaf_categoryUSA	1674.507
	(1710.744)
iaaf_categoryUZB	1597.583
	(2419.239)
iaaf_categoryVEN	1357.899
	(1735.798)
iaaf_categoryVIE	1927.888
	(2024.060)
iaaf_categoryVIN	10 974.743

	(3825.057)
iaaf_categoryZAM	3860.863
	(2962.892)
iaaf_categoryZIM	729.018
	(1997.590)
Num.Obs.	55 512
R2	0.152
R2 Adj.	0.149
AIC	1 061 188.0
BIC	1 062 660.5
Log.Lik.	−530 429.004
F	60.851
RMSE	3416.17

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.