Examining the Impact of Age, Gender, and Environmental Conditions on Marathon Performance

Dingxuan Zhang

October 2024

Abstract

This study investigates the effects of environmental conditions on marathon performance across different ages and genders. Using performance data from five major marathons spanning 15-20 years, we aim to explore how increasing age affects performance, how environmental factors like temperature, humidity, and WBGT influence results, and whether these impacts vary by gender. Key findings are expected to highlight how environmental stress, particularly high WBGT, affects performance, especially in older athletes. This analysis will deepen our understanding of the interaction between age, gender, and environmental factors in marathon running.

Introduction

Marathon running performance is shaped by a range of factors, with environmental conditions like temperature and humidity playing a crucial role. Research shows that performance declines with increasing temperatures, particularly in longer races like marathons, where endurance is key. Older athletes face additional challenges in thermoregulation, making them more vulnerable to heat stress. Gender differences also influence performance, as men and women exhibit varying physiological responses to heat.

This study builds upon previous research by analyzing the performance of male and female marathon runners across a wide age range (14-85 years) under different environmental conditions. The dataset includes detailed weather parameters for five major marathons held over 15-20 years. The analysis focuses on three key aims:

1.Examine effects of increasing age on marathon performance in men and women.

2.Explore the impact of environmental conditions on marathon performance, and whether the impact differs across age and gender.

3.Identify the weather parameters (WBGT, Flag conditions, temperature, etc) that have the largest impact on marathon performance.

Through this analysis, we hope to provide deeper insights into the relationship between environmental factors, age, and gender in marathon performance, contributing to the body of knowledge on endurance exercise in varying weather conditions.

Data Collection

The dataset for this study was obtained from five major marathons held between 1993 and 2016: the Boston Marathon, Chicago Marathon, New York City Marathon, Twin Cities Marathon, and Grandma's Marathon.

These marathons are well-known for their high levels of competition and attract elite runners from around the world. The dataset includes performance data for the fastest male and female runners at each age from 14 to 85 years, recorded as the percentage off the course record for that specific marathon.

In addition to performance data, detailed environmental conditions were recorded for each race. Key weather variables include dry bulb temperature (Td, C), wet bulb temperature (Tw, C), percent relative humidity (%rh), black globe temperature (Tg, C), solar radiation (SR, W/m^2), dew point (DP, C), wind speed (Km/h), and Wet Bulb Globe Temperature (WBGT). WBGT is a weighted average of dry bulb temperature, wet bulb temperature, and black globe temperature, and is commonly used to assess the risk of heat-related illnesses in outdoor activities.

The dataset is critical for addressing the three aims of this study, as it provides both the marathon performance data and corresponding environmental conditions that allow for an in-depth exploration of the impact of age, gender, and weather on marathon performance.

Data Preprocessing

The raw dataset included 14 variables capturing marathon performance, age, gender, and environmental conditions, with an additional dataset containing Air Quality Index (AQI) measurements for different marathons. The data were processed to include 12 variables, focusing on those relevant to this portion of the project. Redundant or irrelevant variables, such as certain marathon-specific metrics, were either dropped or combined into summary variables. The final variables retained for analysis included age, performance metrics (% off course record), weather variables (dry bulb temperature, wet bulb temperature, relative humidity, wind speed, solar radiation), and AQI data.

A few variables required additional cleaning to standardize their values. For instance, some temperature values were formatted incorrectly, requiring conversion from character to numeric. Additionally, inconsistent formatting in the marathon location variable was cleaned by mapping numerical race codes to their corresponding marathon names (e.g., 0 = Boston, 1 = Chicago). These transformations ensured that the data could be accurately merged with the AQI dataset. On top of that, we can see that race and marathon are redundant so we delete race.

The AQI dataset contained multiple records for the same marathon events due to daily AQI measurements. These were aggregated by calculating the mean AQI for each marathon to avoid duplication during the merging process. The datasets were merged based on marathon location, adding AQI as a new environmental variable for further analysis.

Some records contained missing data for key environmental variables. Rows with missing temperature, humidity, wind speed, or AQI values were dropped, as these variables are essential for evaluating the relationship between weather conditions and marathon performance. The final dataset included 11,073 rows after these cleaning steps were completed.

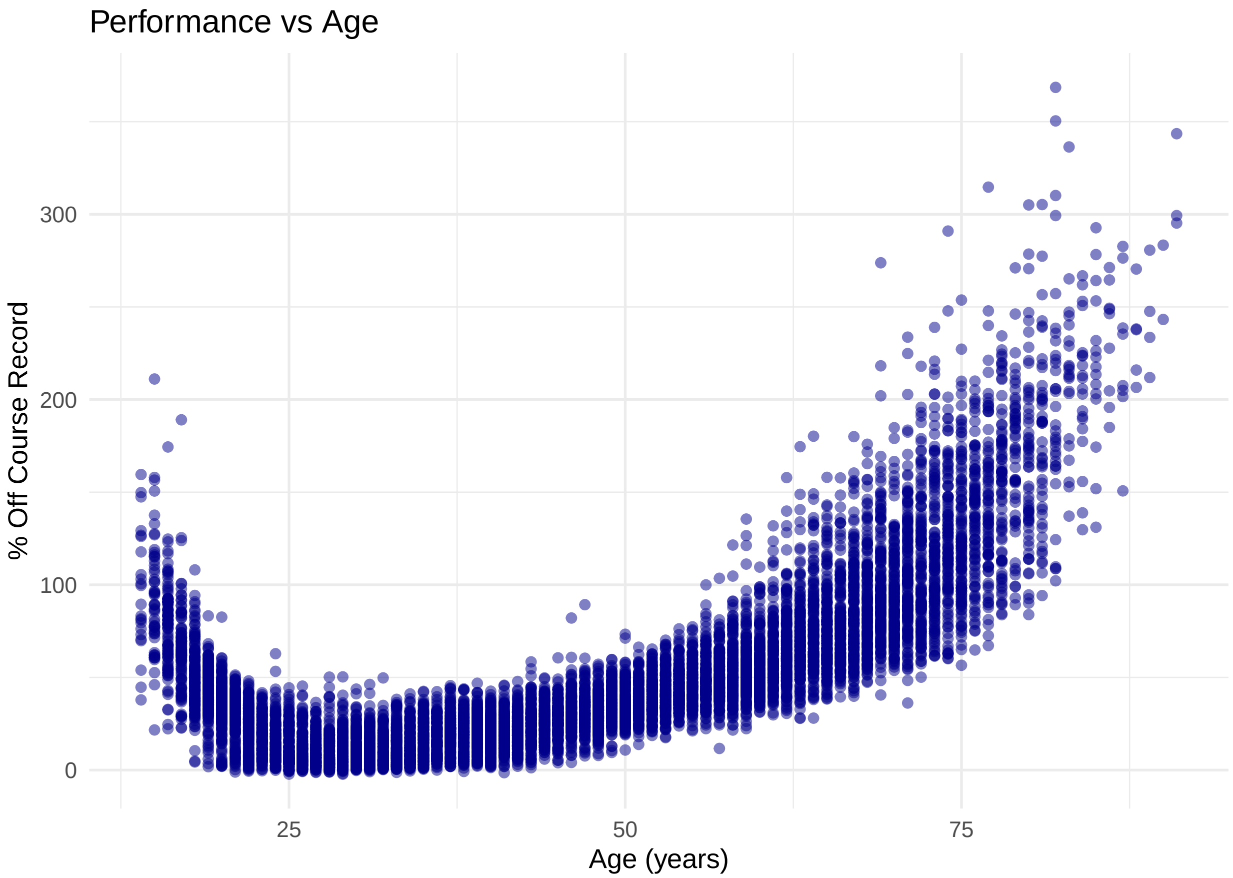
In summary, the cleaned dataset now contains the key variables necessary to explore the effects of age, gender, and environmental conditions on marathon performance. The summary statistics for these variables are presented in Table 1.

Table 1: Table 1: Summary of Marathon Data

No	Variable	Stats / Values	Freqs (% of Valid)	Valid	Missing
1	Year\ [integer]	Mean (sd) : 2006.4 (6)\ min < med < max:\ 1993 < 2006 < 2016\ IQR (CV) : 10 (0)	24 distinct values	11073\ (100.0%)	0\ (0.0%)
2	Sex..0.F..1.M.\ [factor]	1\. Female\ 2\. Male	\5218 (47.1%)\ \5855 (52.9%)	11073\ (100.0%)	0\ (0.0%)
3	Flag\ [character]	1\. Green\ 2\. Red\ 3\. White\ 4\. Yellow	\4706 (42.5%)\ \ 592 (5.3%)\ \3753 (33.9%)\ \2022 (18.3%)	11073\ (100.0%)	0\ (0.0%)
4	Age..yr.\ [integer]	Mean (sd) : 46.5 (18)\ min < med < max:\ 14 < 46 < 91\ IQR (CV) : 30 (0.4)	78 distinct values	11073\ (100.0%)	0\ (0.0%)
5	X.CR\ [numeric]	Mean (sd) : 48.8 (44.7)\ min < med < max:\ -2.3 < 36 < 368.5\ IQR (CV) : 44.2 (0.9)	10663 distinct values	11073\ (100.0%)	0\ (0.0%)
6	Td..C\ [numeric]	Mean (sd) : 13.3 (5.8)\ min < med < max:\ 2 < 12.5 < 28.1\ IQR (CV) : 8.7 (0.4)	79 distinct values	11073\ (100.0%)	0\ (0.0%)
7	Tw..C\ [numeric]	Mean (sd) : 9.4 (5.4)\ min < med < max:\ -1.3 < 8.5 < 21.6\ IQR (CV) : 8.4 (0.6)	92 distinct values	11073\ (100.0%)	0\ (0.0%)
8	X.rh\ [numeric]	Mean (sd) : 42.3 (32)\ min < med < max:\ 0.3 < 53 < 98.3\ IQR (CV) : 63.3 (0.8)	89 distinct values	11073\ (100.0%)	0\ (0.0%)
9	Tg..C\ [numeric]	Mean (sd) : 24.9 (7.7)\ min < med < max:\ 9.5 < 25 < 44.5\ IQR (CV) : 10.5 (0.3)	90 distinct values	11073\ (100.0%)	0\ (0.0%)
10	SR.W.m2\ [numeric]	Mean (sd) : 513.8 (187.6)\ min < med < max:\ 141.4 < 512.7 < 909.5\ IQR (CV) : 254.3 (0.4)	92 distinct values	11073\ (100.0%)	0\ (0.0%)
11	DP\ [numeric]	Mean (sd) : 5.5 (6.9)\ min < med < max:\ -7.4 < 5.1 < 20.3\ IQR (CV) : 10.6 (1.3)	79 distinct values	11073\ (100.0%)	0\ (0.0%)
12	Wind\ [numeric]	Mean (sd) : 9.9 (4.1)\ min < med < max:\ 0 < 10 < 21.8\ IQR (CV) : 4.9 (0.4)	68 distinct values	11073\ (100.0%)	0\ (0.0%)
13	WBGT\ [numeric]	Mean (sd) : 12.9 (5.6)\ min < med < max:\ 1.3 < 12.7 < 25.1\ IQR (CV) : 9.1 (0.4)	92 distinct values	11073\ (100.0%)	0\ (0.0%)
14	marathon\ [character]	1\. Boston\ 2\. Chicago\ 3\. Grandmas\ 4\. NYC\ 5\. Twin Cities	\2088 (18.9%)\ \2427 (21.9%)\ \1884 (17.0%)\ \2799 (25.3%)\ \1875 (16.9%)	11073\ (100.0%)	0\ (0.0%)
15	mean_aqi\ [numeric]	Mean (sd) : 36.6 (4.6)\ min < med < max:\ 28.9 < 35.9 < 42.4\ IQR (CV) : 6.4 (0.1)	28.88!: 1875 (16.9%)\ 33.95!: 1884 (17.0%)\ 35.88!: 2799 (25.3%)\ 40.38!: 2088 (18.9%)\ 42.40!: 2427 (21.9%)\ ! rounded	11073\ (100.0%)	0\ (0.0%)

Aim 1: Examine effects of increasing age on marathon performance in men and women

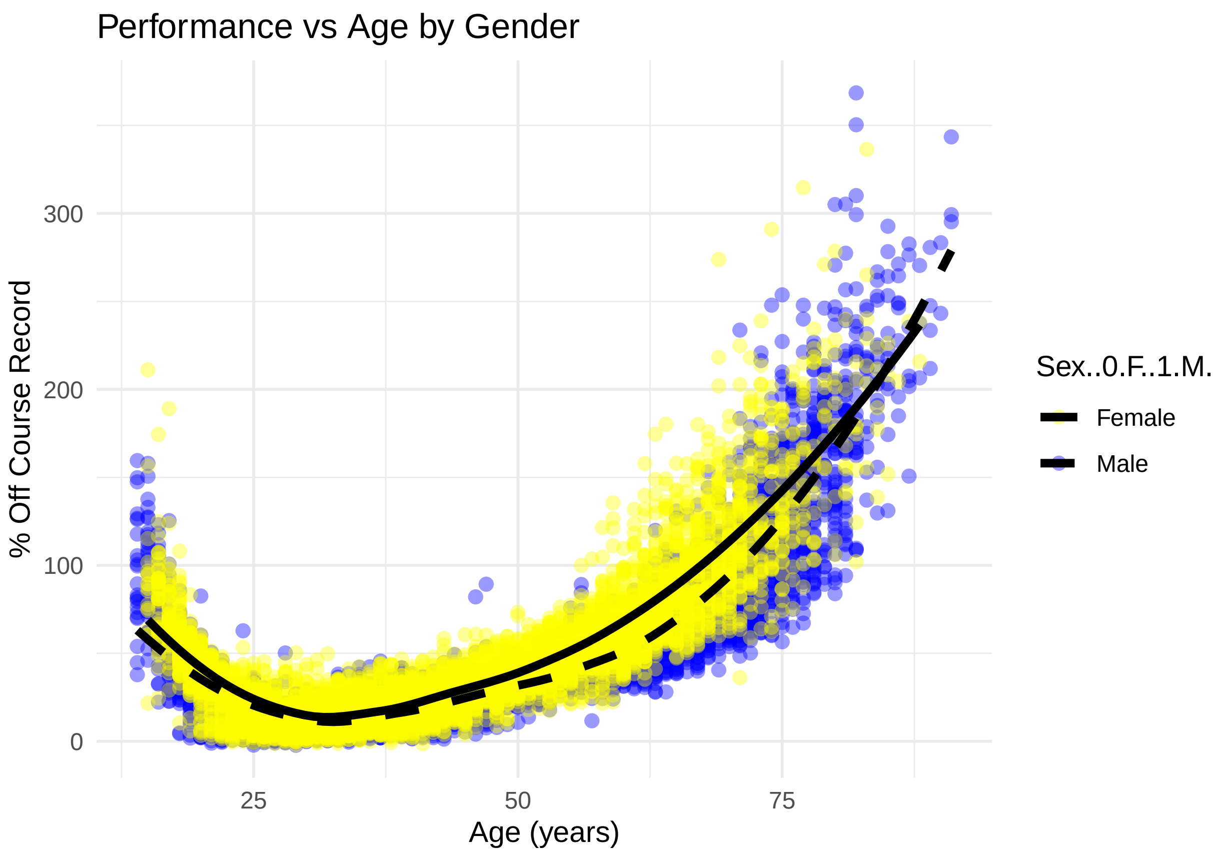
After data pre processing, we can now focus on the proposed three aims to complete the explanatory data analysis. We start from aim 1 which is to examine effects of increasing age on marathon performance in men and women. In this section, we will only be focusing on the three key variables: **Age**, **Performance**, and **Gender**. We shall first provide a basic scatter plot of **age** against **performance**.



The scatter plot shows the relationship between age and marathon performance, measured as the percentage deviation from the course record (%CR). As age increases, there is a clear trend where performance tends to worsen, with older runners performing further from the course record. While there is variation in performance at all ages, younger participants generally perform better, staying closer to the course record, whereas older runners experience greater deviations. Interestingly, some younger runners also perform further from the course record, indicating that factors beyond age may influence performance. This plot provides an initial understanding of how performance changes with age, setting the foundation for further analysis, including the role of gender.

Building on the initial insights from the scatter plot, we now turn to examining how the relationship between age and performance varies by gender. Stratifying the data by gender will allow us to explore whether men and women experience similar performance declines as they age, or if there are notable differences between the two groups.

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



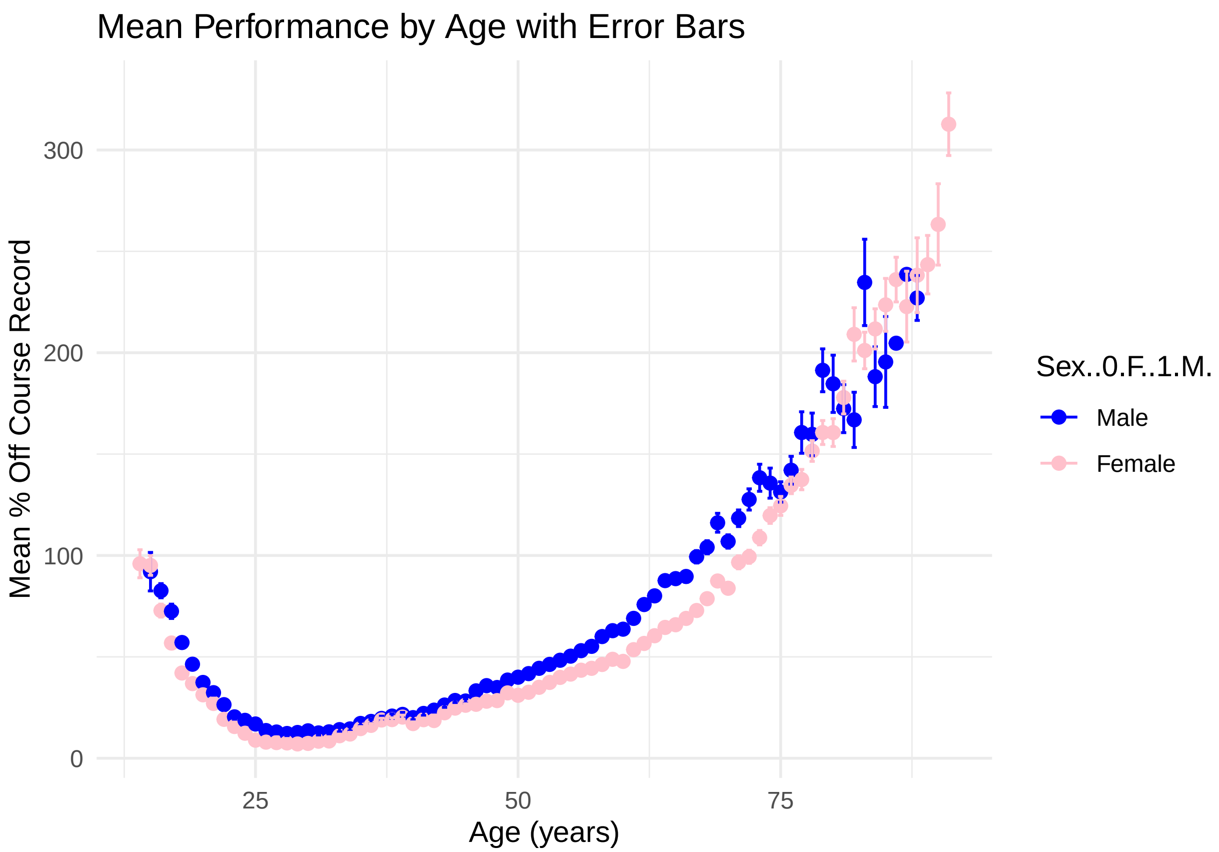
This plot shows the relationship between age and marathon performance (% off course record) for both men and women, with each gender represented by different colors and line types. The points display individual performance data, with females in yellow and males in blue. The smoother lines, drawn in red, provide a clearer visualization of the overall trend for each gender, with a solid line for females and a dashed line for males.

From the plot, it's evident that both men and women experience a non-linear relationship between age and performance. Among younger runners, performance initially improves with age, reaching a peak in the late 20s or early 30s. After this peak, performance gradually declines as age increases, with the trend being more pronounced in older athletes. This pattern is consistent across both genders, but the trend line for men suggests a slightly steeper decline in performance at older ages compared to women, who show a more gradual decline.

This non-linear relationship highlights that while both genders exhibit age-related declines in performance, the nature and rate of decline differ slightly. The smoother lines help us capture the peak performance age and the subsequent decrease with age, offering a clearer understanding of how age affects marathon performance across both men and women.

We then use a standard error plot to examine how the variance changes with ages.

```
## `summarise()` has grouped output by 'Age..yr.'. You can override using the
## `.groups` argument.
```



This plot displays the mean marathon performance (% off course record) at different ages for both men and women, with vertical error bars representing the standard error of the mean. The points indicate the average performance for each age group, and the error bars give an indication of the variance or uncertainty around these averages.

From the plot, we observe that the larger error bars for both men and women at older ages suggest that there is greater variability in performance among older runners. This could indicate that while some older runners maintain strong performances, others experience more significant declines, leading to higher variability within this age group. In contrast, the smaller error bars among younger runners suggest that their performance is more consistent, with less variation around the mean.

summary for aim 1

The analysis of age effects on marathon performance revealed a clear non-linear relationship between age and performance for both men and women. Performance generally improves as runners reach their late 20s to early 30s, after which a decline is observed with increasing age. While both genders experience this pattern, the rate and consistency of the decline differ. Men exhibit a decline in performance with greater variability at older ages, whereas women show a more gradual decline with relatively more consistent performance. These findings suggest that both age and gender play significant roles in determining marathon performance, with different trajectories observed for men and women across the lifespan.

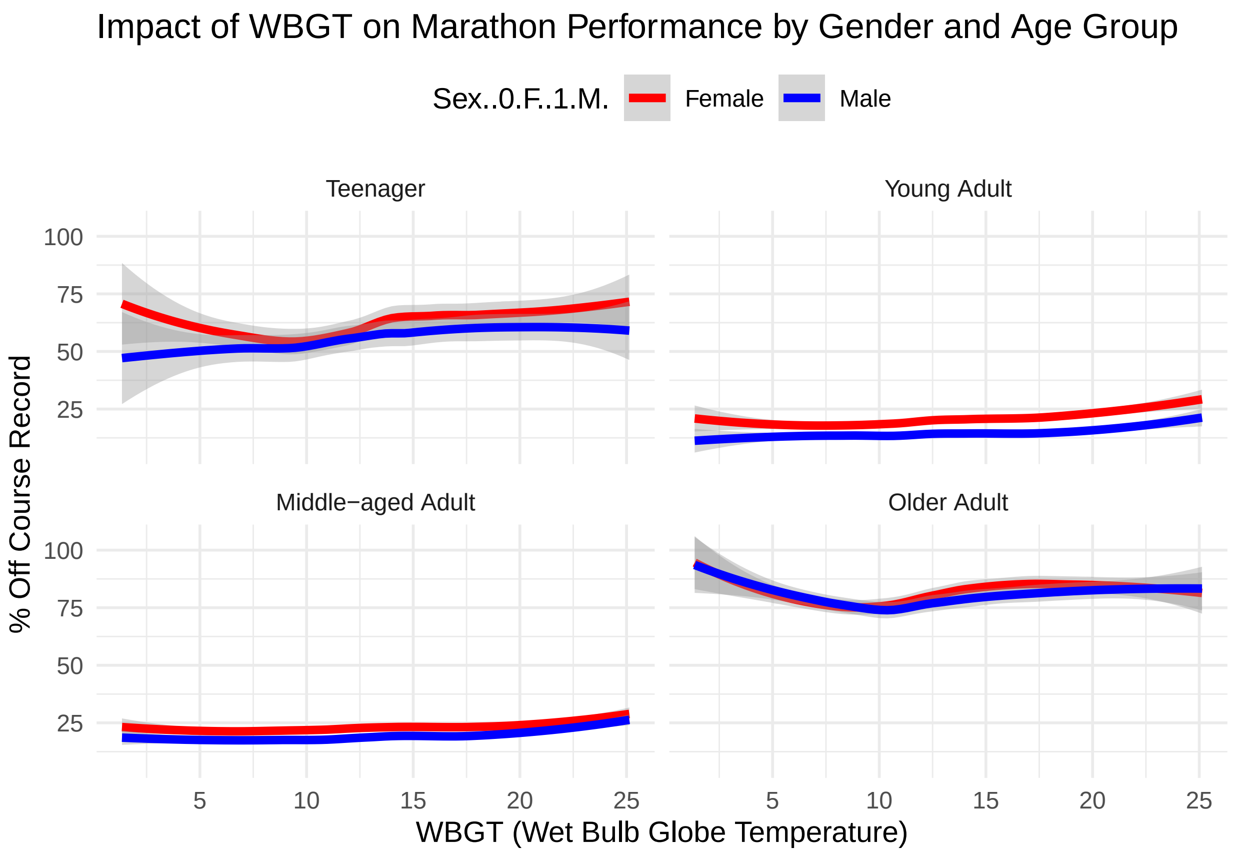
Aim 2: Explore the impact of environmental conditions on marathon performance, and whether the impact differs across age and gender.

This section investigates the relationship between various **environmental conditions** and marathon performance, focusing on whether these effects vary across different **genders** and **age groups**. Specifically, we examined the impact of environmental factors such as **Wet Bulb Globe Temperature (WBGT)** , **wind**, and **air quality index (AQI)**. To explore these relationships, we generated visualizations that depict how each environmental condition influences performance, measured as the percentage deviation from the course record (%CR).

The initial analysis focused on WBGT, which serves as an overall measure of heat stress by combining temperature, humidity, and solar radiation, and therefore we only evaluate WBGT, wind, and AQI here instead of all environmental conditions. The smooth trend lines, stratified by gender, highlight the impact of WBGT on marathon performance. Higher WBGT values are associated with worsening performance for both men and women, indicating the detrimental effect of increased heat stress. The decline in performance, however, appears more pronounced for men, suggesting a potential difference in gender-specific responses to environmental heat stress.

To better understand these interactions, participants were categorized into **four distinct age groups**: **Teenagers (14-19 years)**, **Young Adults (20-29 years)**, **Middle-aged Adults (30-49 years)**, and **Older Adults (50+ years)**.

The figure below displays the relationship between WBGT and marathon performance, measured as the percentage deviation from the course record (%CR), for each age group and stratified by gender. The smooth trend lines, along with the confidence intervals, help us visualize the impact of WBGT across the different groups.

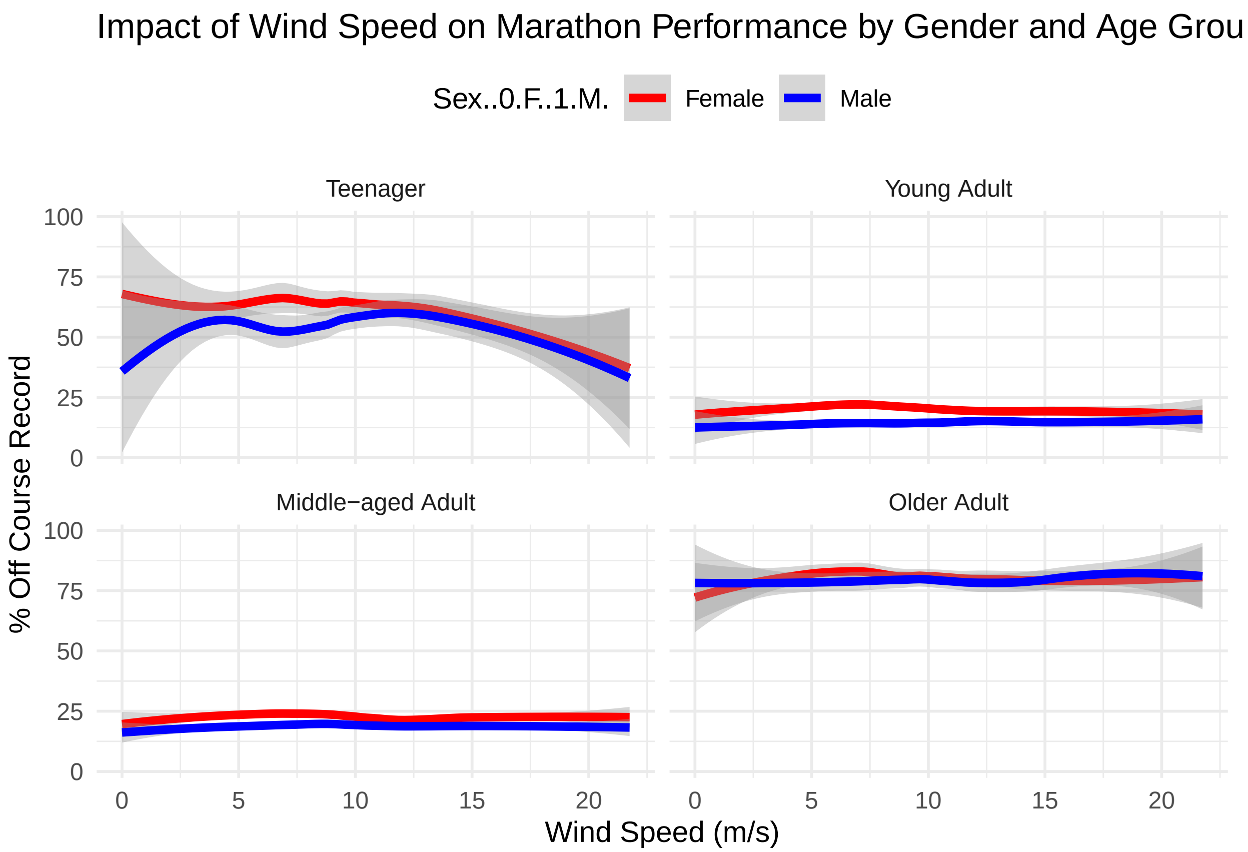


The figure reveals that increasing WBGT generally leads to worsened performance across all age groups, as evidenced by higher deviations from the course record (%CR). For teenagers (14-19 years), performance appears relatively stable with increasing WBGT, indicating potential resilience to heat stress. However, this age group also shows wide confidence intervals, suggesting variability in individual responses.

Young adults (20-29 years) experience a slight decline in performance as WBGT increases, but the effect is less severe compared to older age groups. This may indicate a peak capacity to handle environmental stress among runners in their 20s. Middle-aged (30-49 years) and older adults (50+ years), however, show steeper declines in performance with increasing WBGT, highlighting their increased vulnerability to heat stress. The confidence intervals are also wider for these groups, particularly for older adults, suggesting greater variability in performance under heat stress.

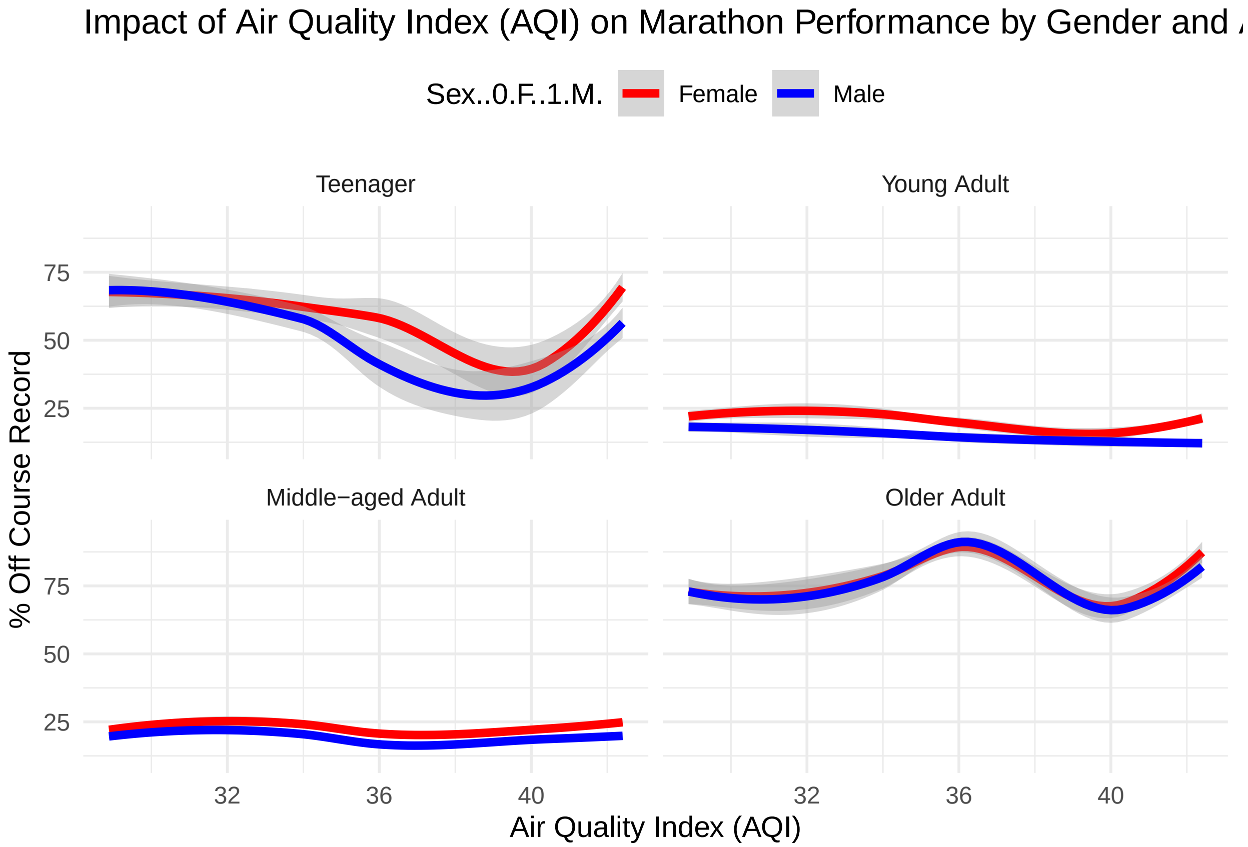
Gender differences are evident across all age groups. Men show a sharper decline in performance compared to women, particularly in the middle-aged and older adult categories. This indicates that men may be more sensitive to the effects of heat stress as they age. Women, on the other hand, appear to maintain a more gradual decline, potentially indicating better heat adaptation or endurance under increasing WBGT conditions.

We now proceed to wind speed.



The effect of wind speed on marathon performance varies across age groups and genders. For teenagers and young adults, moderate wind speeds seem to enhance performance by aiding cooling, but higher wind speeds cause slight declines due to increased resistance. Middle-aged and older adults are more negatively impacted by stronger winds, likely due to reduced physical capacity to manage resistance.From gender perspective, men are generally more negatively affected by strong winds compared to women, especially in older age groups, possibly reflecting differences in pacing strategies.

We now proceed to AQI.



The impact of AQI on marathon performance shows a general decline in performance as AQI increases, indicating that higher pollution levels negatively affect endurance running. Older adults (50+ years) are particularly sensitive to poorer air quality, showing the most significant decline in performance with rising AQI. Middle-aged and young adults also experience declines, but the effect is less severe compared to older athletes. Teenagers seem less affected by AQI, suggesting that their performance may not yet be as influenced by air quality compared to older runners. Gender differences are apparent, with men generally experiencing a greater performance decline with increasing pollution compared to women, particularly in older age groups. This may suggest that women are better at mitigating the effects of pollution, potentially due to differences in pacing or physiological responses.

sumamry of aim 2

The analysis of WBGT (Wet Bulb Globe Temperature), wind speed, and air quality (AQI) on marathon performance revealed distinct patterns across age groups and genders.

WBGT was found to have a negative impact on performance, particularly for older adults (50+ years) and middle-aged adults (30-49 years). Men experienced a steeper decline in performance with rising WBGT compared to women, suggesting gender differences in heat tolerance.

Wind speed showed a dual effect: moderate wind had a beneficial impact on performance, especially for younger runners, likely due to its cooling effects. However, high wind speeds negatively affected performance, particularly for older adults and men, who showed greater sensitivity to wind resistance.

Air quality (AQI) had a similarly negative effect, with older adults displaying the most significant performance decline in polluted conditions. Men were generally more impacted by poor air quality than women, highlighting a gender difference in the response to pollution.

Aim 3: Identify the weather parameters (WBGT, Flag conditions, temperature, etc) that have the largest impact on marathon performance.

In this section, we will be using statistical tests to identify the weather parameters having most significant impact on marathon performance. We first apply linear regression to explore the impact of each environmental factor in a straightforward manner. Linear regression allows us to understand the direction and magnitude of each predictor's effect while controlling for other variables. However, due to the potential non-linear interactions we already found in some previous research on age, we also employ a random forest model to capture more complex relationships that linear regression might miss. This approach helps us validate the findings and provides a more comprehensive understanding of the key predictors.

Linear regresssion model

The initial analysis uses a multiple regression model to quantify the relative effect of each environmental condition on marathon performance. This model allows us to estimate regression coefficients, which provide a clear indication of the significance of each predictor's relationship with performance.

The multiple regression model provides an understanding of how each environmental factor impacts marathon performance while controlling for age and gender. Including age and sex in the model helps to control for their effects and ensures that the analysis accurately captures the influence of environmental conditions on performance, reducing the risk of confounding. By analyzing the significance of each coefficient, we can determine which factors have a statistically significant impact on performance. Based on the regression summary, WBGT, dry bulb temperature (Td..C), and age were found to be statistically significant predictors of marathon performance, indicating that these factors play an important role in determining performance outcomes. Relative humidity (X.rh), wind speed (Wind), and air quality (mean_aqi) were not statistically significant in this model, suggesting their effects may be less significant or influenced by other interactions not captured by the linear model.

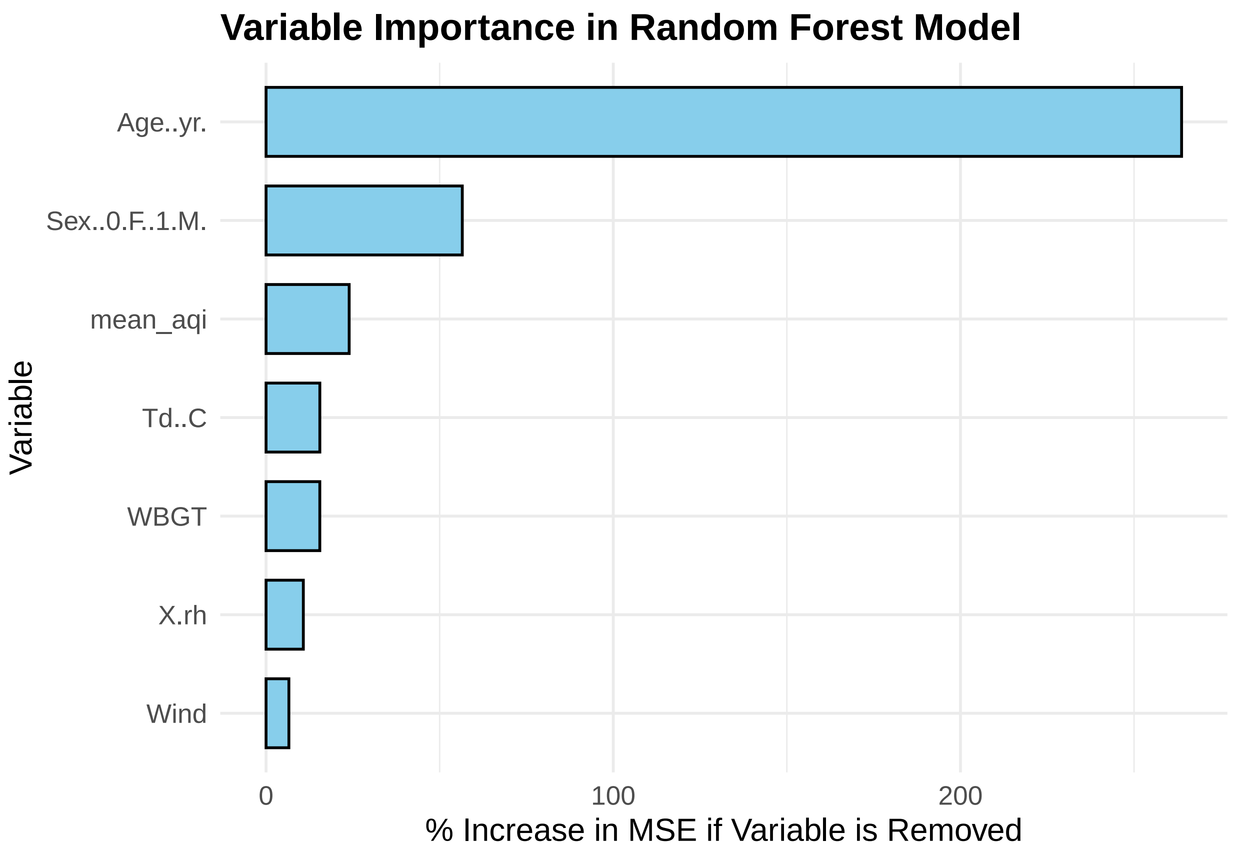
The linear regression approach offers a straightforward way to determine which environmental factors have the most substantial effects while holding other factors constant. However, this method has limitations, especially in its ability to capture non-linear relationships or interactions between variables that may be present in a complex real-world scenario like marathon performance.

```
##
## Call:
## lm(formula = X.CR ~ WBGT + Td..C + X.rh + Wind + mean_aqi + Age..yr. +
##     Sex..0.F..1.M., data = merged_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -57.770 -19.423  -9.797   7.457 261.214
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -32.42597    2.99510 -10.826  < 2e-16 ***
## WBGT                -0.59181    0.26985  -2.193   0.0283 *
## Td..C                0.99425    0.25477   3.902 9.58e-05 ***
## X.rh                 0.01077    0.01017   1.059   0.2896
## Wind                -0.13358    0.08452  -1.581   0.1140
## mean_aqi            -0.07454    0.06719  -1.109   0.2673
## Age..yr.             1.75944    0.01687 104.309  < 2e-16 ***
## Sex..0.F..1.M.Male  -4.79922    0.60587  -7.921 2.58e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 31.72 on 11065 degrees of freedom
## Multiple R-squared:  0.4973, Adjusted R-squared:  0.4969
## F-statistic:  1563 on 7 and 11065 DF,  p-value: < 2.2e-16
```

Random Forest model

To capture potential non-linear effects and interactions among the environmental predictors, we use a random forest model. Unlike linear regression, which assumes a linear relationship between predictors and the outcome, random forest is a non-parametric ensemble method that can handle complex interactions without requiring strict assumptions about the data structure.

The variable importance plot generated by the random forest model highlights which environmental conditions have the greatest influence on marathon performance. Among the environmental factors, mean AQI (Air Quality Index) was identified as the most important predictor, suggesting that air pollution significantly impacts marathon performance. WBGT (Wet Bulb Globe Temperature) and temperature (Td..C) also showed notable importance, indicating that heat stress and temperature extremes are crucial factors in determining performance outcomes. Temperature (Td..C) also shows notable importance, which aligns with the understanding that extreme temperatures—whether high or low—affect endurance running outcomes. Wind speed (Wind) has a moderate influence, suggesting that wind conditions can either help or hinder performance depending on the circumstances. Relative humidity (X.rh) and air quality index (mean_aqi) have lower importance compared to other factors, indicating that their effects may be less critical or more context-dependent in relation to marathon performance. Overall, this analysis underscores the significant influence of air quality and heat-related environmental factors, such as mean AQI, WBGT, and temperature, on marathon outcomes, highlighting the importance of considering these conditions to optimize performance and safety for athletes., which is crucial for understanding how to optimize performance and safety for athletes in different weather conditions.



summary of aim 3

Based on both regression analysis and random forest modeling, the key environmental predictors of marathon performance were identified as mean AQI, WBGT, temperature, and wind speed. Among these, mean AQI was found to have the highest importance in the random forest model, suggesting that air quality plays a crucial role in determining performance outcomes, especially under polluted conditions. Among these, temperature was found to be most statistically significant in linear model. These findings highlight the significant influence of environmental factors on marathon outcomes.

Conclusion

The analysis aimed to investigate the effects of age, environmental conditions, and specific weather parameters on marathon performance, focusing on three key aims. Aim 1 examined the effect of age on marathon performance, revealing that performance tends to decline with increasing age, with younger runners performing better. The smooth lines showed clear trends, emphasizing that age is a significant factor in determining an athlete's ability to achieve optimal performance.

Aim 2 explored the impact of environmental conditions such as WBGT, wind speed, and AQI on performance, and whether these effects differed by age and gender. The results showed that WBGT and wind speed significantly influenced performance, with higher WBGT leading to worse outcomes, while wind speed had mixed effects. Additionally, differences in performance across age groups and genders were evident, highlighting the importance of environmental adaptations.

Aim 3 focused on identifying the key environmental predictors of marathon performance. Both linear regression and random forest modeling were used, and the random forest analysis revealed that mean AQI was the most critical predictor, followed by WBGT and temperature. These findings suggest that air quality and heat-related conditions are the most influential environmental factors affecting marathon performance.

Overall, this comprehensive analysis demonstrates that both age-related factors and environmental conditions are critical determinants of marathon performance. Age plays a major role in how athletes perform, while environmental factors such as air quality, temperature, and wind significantly affect performance outcomes. These findings underscore the importance of optimizing training and preparation strategies by considering both age and environmental conditions to enhance performance and reduce risks for marathon runners. Additionally, heat-related factors such as WBGT and temperature also substantially impact performance, highlighting the need to prepare for temperature variations during marathons. Wind speed showed mixed effects, either aiding or hindering performance depending on its intensity. These results emphasize the importance of considering both environmental and demographic factors to optimize marathon performance and ensure athlete safety in various weather conditions.. Among these, WBGT was found to be statistically significant and consistently showed an impact on performance. Temperature also had a substantial impact, with extreme values affecting runners' outcomes. Wind speed had mixed effects, with moderate wind aiding performance and high wind acting as a hindrance. Relative humidity and AQI were less impactful compared to other factors. These findings highlight the significant influence of environmental factors on marathon outcomes, emphasizing the need for tailored preparation strategies to mitigate the effects of challenging weather conditions.

References

1. Ely, B. R., Cheuvront, S. N., Kenefick, R. W., & Sawka, M. N. (2010). Aerobic performance is degraded, despite modest hyperthermia, in hot environments. Med Sci Sports Exerc, 42(1), 135-41.

2. Ely, M. R., Cheuvront, S. N., Roberts, W. O., & Montain, S. J. (2007). Impact of weather on marathon-running performance. Medicine and science in sports and exercise, 39(3), 487-493.

3. Kenney, W. L., & Munce, T. A. (2003). Invited review: aging and human temperature regulation. Journal of applied physiology, 95(6), 2598-2603.

4. Besson, T., Macchi, R., Rossi, J., Morio, C. Y., Kunimasa, Y., Nicol, C., ... & Millet, G. Y. (2022). Sex differences in endurance running. Sports medicine, 52(6), 1235-1257.

5. Yanovich, R., Ketko, I., & Charkoudian, N. (2020). Sex differences in human thermoregulation: relevance for 2020 and beyond. Physiology, 35(3), 177-184.

Code Appendix:

```
knitr::opts_chunk$set(echo = FALSE)
library(tidyverse)
library(summarytools)
library(kableExtra)
library(RColorBrewer)
library(ggplot2)
library(dplyr)
library(randomForest)

# data cleaning and merging

# Load the marathon dataset
marathon_data <- read.csv("project1.csv")

# Clean the dataset by removing rows with missing environmental data
marathon_cleaned <- marathon_data %>%
  drop_na(`Td..C`, `Tw..C`, `X.rh`, `Tg..C`, `SR.W.m2`, `DP`, `Wind`, `WBGT`) %>%
  mutate(
    `Td..C` = as.numeric(`Td..C`),
    `Tw..C` = as.numeric(`Tw..C`),
    `X.rh` = as.numeric(`X.rh`),
    `Tg..C` = as.numeric(`Tg..C`),
    `SR.W.m2` = as.numeric(`SR.W.m2`),
    DP = as.numeric(DP),
    Wind = as.numeric(Wind),
    WBGT = as.numeric(WBGT)
  )

# Load the AQI dataset
aqi_data <- read.csv("aqi_values.csv")

# Aggregate the AQI data to avoid duplicates (e.g., take the mean AQI for each marathon)
# Assuming there might be multiple AQI entries per marathon, we take the mean of 'aqi'
aqi_aggregated <- aqi_data %>%
  group_by(marathon) %>%
  summarize(mean_aqi = mean(aqi, na.rm = TRUE))

# Create a mapping between race codes and marathon names
race_mapping <- c("Boston", "Chicago", "NYC", "Twin Cities", "Grandmas")
marathon_cleaned <- marathon_cleaned %>%
  mutate(marathon = factor(`Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D.`, labels = race_mapping))

# Merge the marathon dataset with the aggregated AQI data on marathon name
merged_data <- left_join(marathon_cleaned, aqi_aggregated, by = "marathon")

# Fix the data type of sex and delete redundant race
merged_data <- merged_data %>%
  mutate(Sex..0.F..1.M. = factor(Sex..0.F..1.M., levels = c(0, 1), labels = c("Female", "Male")))
merged_data <- merged_data %>%
  select(-`Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D.`)
```

```
# create table
summary_table <- dfSummary(merged_data,
                           plain.ascii = TRUE,  # Use ASCII format for PDF compatibility
                           style = "grid",
                           varlabels = c(
                             "Marathon" = "marathon",  # Use the 'marathon' column for race names
                             "Sex" = "Sex..0.F..1.M.",
                             "Age (years)" = "Age..yr.",
                             "%CR" = "X.CR",
                             "Dry Bulb Temp (°C)" = "Td..C",
                             "Wet Bulb Temp (°C)" = "Tw..C",
                             "Rel Humidity (%)" = "X.rh",
                             "Globe Temp (°C)" = "Tg..C",
                             "Solar Rad (W/m2)" = "SR.W.m2",
                             "Dew Point" = "DP",
                             "Wind Speed" = "Wind",
                             "WBGT" = "WBGT",
                             "Mean AQI" = "mean_aqi"
                           ),
                           headings = FALSE,
                           graph.col = FALSE)

# Convert the summary table to a data frame for further customization if necessary
df_table <- as.data.frame(summary_table)

# Print the summary table using kable for LaTeX/PDF output
kable(df_table, "latex", booktabs = TRUE, caption = "Table 1: Summary of Marathon Data") %>%
  kable_styling(latex_options = c("scale_down", "hold_position"))  # Adjust table size to fit page
# Basic scatter plot of performance (%CR) vs age
ggplot(merged_data, aes(x = Age..yr., y = X.CR)) +
  geom_point(alpha = 0.5, color = "darkblue") +
  labs(title = "Performance vs Age", x = "Age (years)", y = "% Off Course Record") +
  theme_minimal()

# Plot with line drawn on top of points and distinct colors
ggplot(merged_data, aes(x = Age..yr., y = X.CR)) +
  geom_point(aes(color = Sex..0.F..1.M.), alpha = 0.4, size = 2) +  # Points with gender-specific color
  geom_smooth(aes(linetype = Sex..0.F..1.M.), method = "loess", se = FALSE, size = 1.5, color = "black"
  labs(title = "Performance vs Age by Gender", x = "Age (years)", y = "% Off Course Record") +
  scale_color_manual(values = c("yellow", "blue"), labels = c("Female", "Male")) +  # Colors for points
  scale_linetype_manual(values = c("solid", "dashed"), labels = c("Female", "Male")) +  # Different lin
  theme_minimal()

# Summarize the data by calculating mean performance and standard error by age and gender
summary_data <- merged_data %>%
  group_by(Age..yr., Sex..0.F..1.M.) %>%
  summarise(
    mean_performance = mean(X.CR, na.rm = TRUE),
    se_performance = sd(X.CR, na.rm = TRUE) / sqrt(n())  # Standard Error
  )

# Plot with error bars (without smooth line)
ggplot(summary_data, aes(x = Age..yr., y = mean_performance, color = Sex..0.F..1.M.)) +
  geom_point(size = 2) +  # Plot points for mean performance
  geom_errorbar(aes(ymin = mean_performance - se_performance, ymax = mean_performance + se_performance)
  labs(title = "Mean Performance by Age with Error Bars",
       x = "Age (years)",
       y = "Mean % Off Course Record") +
  scale_color_manual(values = c("blue", "pink"), labels = c("Male", "Female")) +
  theme_minimal()


# Create new age groups
merged_data$AgeGroup <- cut(merged_data$Age..yr.,
                            breaks = c(13, 19, 29, 49, Inf),  # Adding Inf to include any values
                            labels = c("Teenager", "Young Adult", "Middle-aged Adult", "Older Adult"),
                            include.lowest = TRUE)

# Plot WBGT vs Performance, stratified by gender and age group
ggplot(merged_data, aes(x = WBGT, y = X.CR, color = Sex..0.F..1.M.)) +
  geom_smooth(method = "loess", se = TRUE, size = 1.5) +  # Add smooth trend lines with confidence inte
  facet_wrap(~ AgeGroup) +  # Create separate panels for each age group
  labs(title = "Impact of WBGT on Marathon Performance by Gender and Age Group",
       x = "WBGT (Wet Bulb Globe Temperature)",
       y = "% Off Course Record") +
  scale_color_manual(values = c("red", "blue"), labels = c("Female", "Male")) +  # Color points by gend
  theme_minimal() +
  theme(legend.position = "top")
# Plot Wind Speed vs Performance, stratified by gender and updated age groups
ggplot(merged_data, aes(x = Wind, y = X.CR, color = Sex..0.F..1.M.)) +
  geom_smooth(method = "loess", se = TRUE, size = 1.5) +  # Add smooth trend lines with confidence inte
  facet_wrap(~ AgeGroup) +  # Create separate panels for each age group
  labs(title = "Impact of Wind Speed on Marathon Performance by Gender and Age Group",
       x = "Wind Speed (m/s)",
       y = "% Off Course Record") +
  scale_color_manual(values = c("red", "blue"), labels = c("Female", "Male")) +
  theme_minimal() +
  theme(legend.position = "top")

# Plot AQI vs Performance, stratified by gender and updated age groups
ggplot(merged_data, aes(x = mean_aqi, y = X.CR, color = Sex..0.F..1.M.)) +
  geom_smooth(method = "loess", se = TRUE, size = 1.5) +  # Add smooth trend lines with confidence inte
  facet_wrap(~ AgeGroup) +  # Create separate panels for each age group
  labs(title = "Impact of Air Quality Index (AQI) on Marathon Performance by Gender and Age Group",
       x = "Air Quality Index (AQI)",
       y = "% Off Course Record") +
  scale_color_manual(values = c("red", "blue"), labels = c("Female", "Male")) +
  theme_minimal() +
  theme(legend.position = "top")

# Run a linear model to identify key environmental predictors of performance
lm_model <- lm(X.CR ~ WBGT + Td..C + X.rh + Wind + mean_aqi + Age..yr. + Sex..0.F..1.M., data = merged_

# Summary of the model to interpret coefficients
summary(lm_model)
# Train a random forest model to assess variable importance
```

```r
set.seed(0)  # Set seed for reproducibility
rf_model <- randomForest(X.CR ~ WBGT + Td..C + X.rh + Wind + mean_aqi + Age..yr. + Sex..0.F..1.M.,
                         data = merged_data,
                         ntree = 200,          # Increase number of trees to 200 for stability
                         mtry = 3,              # Set number of variables tried at each split to 3
                         importance = TRUE)

# Convert variable importance to a data frame
var_importance <- as.data.frame(importance(rf_model))
var_importance$Variable <- rownames(var_importance)

# Create a ggplot for variable importance
ggplot(var_importance, aes(x = reorder(Variable, `%IncMSE`), y = `%IncMSE`)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black", width = 0.7) +
  coord_flip() +
  labs(title = "Variable Importance in Random Forest Model",
       x = "Variable",
       y = "% Increase in MSE if Variable is Removed") +
  theme_minimal() +
  theme(axis.text = element_text(size = 10),
        axis.title = element_text(size = 12),
        plot.title = element_text(size = 14, face = "bold"))
```