Explore the impact of environmental conditions on marathon performance in gender

Jing Fu

October 2024

Abstract

This study investigates the impact of age, gender, and environmental conditions on marathon performance, aiming to identify key weather parameters that most significantly affect outcomes. Using data from major marathons, we examined the effects of increasing age on performance across both sexes, explored how environmental factors like WBGT, temperature, and humidity influence performance, and determined whether these effects vary by age and gender. Our results highlight that extreme temperature conditions, particularly high WBGT levels and dry bulb temperature (Td), significantly impair marathon performance. Solar radiation (SR) and dew point temperature (DP) showed mixed effects, with SR sometimes improving performance under moderate conditions. However, limitations include the dataset's representativeness and the lack of individual-level data such as training status. These findings underscore the need for careful race planning and athlete preparation under varying environmental conditions.

Introduction

Endurance sports such as marathon running are profoundly influenced by physiological and environmental factors. As global temperatures rise, understanding the impact of environmental conditions on athletic performance, especially in endurance sports, becomes increasingly crucial. Previous research has demonstrated that aerobic performance degrades in hot environments, with even modest hyperthermia reducing endurance capacity significantly[1]. Moreover, the interaction between age, gender, and environmental stressors adds layers of complexity to this issue. Older adults, for example, face greater thermoregulatory challenges, which impair heat dissipation and exacerbate the impact of high temperatures[3]. Sex differences also play a crucial role in how athletes respond to environmental stressors. Males and females exhibit significant differences in endurance performance and thermoregulation mechanisms, which are crucial to understanding differential performance in marathon races[4]. Ely et al. (2007) have underscored the significance of environmental conditions like temperature and humidity on marathon performance, yet gaps remain in our understanding of how these factors interact across different age groups and between genders.

Given this backdrop, our research aims to dissect the complex interplay of age, gender, and environmental conditions on marathon performance. Specifically, we aim to: 1) Examine the effects of increasing age on marathon performance in both men and women. 2) Explore the impact of environmental conditions, such as temperature and humidity, on marathon performance and investigate whether these effects vary across different age groups and between genders. 3) Identify key weather parameters that most significantly affect marathon performance.

This study utilized a dataset that included the best single age marathon results for males and females aged 14 to 91 from 1993 to 2016, as well as detailed environmental conditions, to investigate the above three questions.

Data Collection

Marathon race results were sourced from five major marathons: Boston, New York, Twin Cities, Grandma's, and Chicago, covering the years 1993 to 2016. The dataset includes top single-age performances for both male and female runners aged 14 to 91, capturing individual performance metrics and detailed environmental conditions. Each performance is linked to a unique race code, ranging from 0 (Boston) to 4 (Grandma's).

Performance Evaluation Runner performances were evaluated by comparing their finishing times to the current course records for their respective age and gender categories. The deviation from the record is calculated as:(Finishing Time - Course Record)/Course Record $\times$ 100. This percentage reflects performance relative to the record, adjusting for any annual improvements in conditions or athletic capability, ensuring consistent cross-year comparisons.

Environmental Conditions Collected environmental data includes dry bulb temperature (Td), wet bulb temperature (Tw), relative humidity (%rh), black globe temperature (Tg), solar radiation (SR, W/m^2), dew point (DP,), and wind speed (Wind, km/hr). These measures are critical for determining the heat stress, quantified by the Wet Bulb Globe Temperature (WBGT). WBGT integrates these factors into a comprehensive index of heat stress experienced by runners. Risk flags based on WBGT categorize heat illness potential from low (White Flag, WBGT <10) to extreme (Black Flag, WBGT >28).

All variables will be considered to explore the three research questions mentioned above.

Data Preprocessing

In this study, we conducted data preprocessing on the dataset including 11,564 records with 14 variables.

Categorical variables in the dataset were transformed into factors to facilitate subsequent analyses. Specifically: The Race variable was converted into a factor with levels ranging from 0 to 4, each representing a different marathon. The Year variable was also transformed into a factor with levels from 1993 to 2016. The Sex variable was categorized into two levels: 0 (female) and 1 (male). The Flag variable was transformed into a factor representing different risk levels based on WBGT, including White, Green, Yellow, Red, and Black.

We identified 491 missing values for variable Flag, Td, Tw, rh, Tg, SR, DP, Wind, WBGT. The missing values were primarily concentrated in specific years (notably 2011 and 2012) and specific races (1,2,3,4), indicating potential systemic issues in data collection. Given the critical role of environmental variables in our study, we opted to remove records containing any missing values. This decision was due to the lack of information to impute. This process reduced the dataset to 11,073 records. Missing patterns are shown in Figure 1a and 1b. The summary are reported in Table 1.

According to Table 1, it can be observed that there are there is no race for WBGT $>$ 28. And under all conditions, the proportion of men and women participating in the competition is the same. And it can be preliminarily assumed that as the temperature rises, the performance of runners decreases.

Figure 1a. NA in Variables and Race

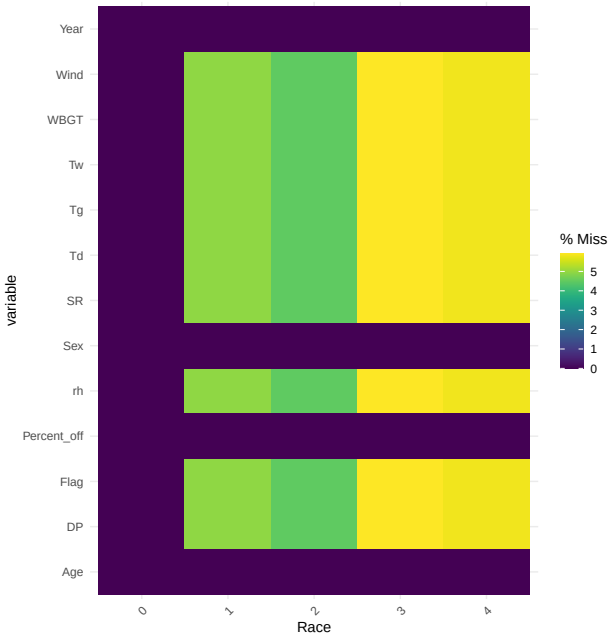


Figure 1b. NA in Variables and Year

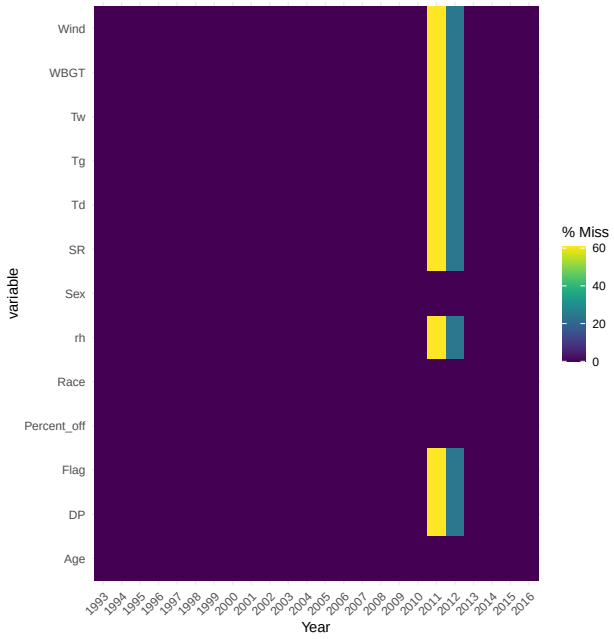


Table 1: Summary Statistics of Marathon Performance by Flag

Characteristic	**White** N = 3,753	**Green** N = 4,706	**Yellow** N = 2,022	**Red** N = 592
Race				
Boston	1,040 (28%)	810 (17%)	115 (5.7%)	123 (21%)
Chicago	732 (20%)	1,459 (31%)	120 (5.9%)	116 (20%)
NYC	1,394 (37%)	901 (19%)	504 (25%)	0 (0%)
Twin Cities	587 (16%)	834 (18%)	338 (17%)	116 (20%)
Grandma's	0 (0%)	702 (15%)	945 (47%)	237 (40%)
Year				
1993	0 (0%)	0 (0%)	125 (6.2%)	0 (0%)
1994	0 (0%)	0 (0%)	127 (6.3%)	0 (0%)
1995	124 (3.3%)	0 (0%)	0 (0%)	0 (0%)
1996	124 (3.3%)	112 (2.4%)	0 (0%)	0 (0%)
1997	0 (0%)	114 (2.4%)	128 (6.3%)	0 (0%)
1998	0 (0%)	338 (7.2%)	0 (0%)	0 (0%)
1999	249 (6.6%)	0 (0%)	0 (0%)	0 (0%)
2000	235 (6.3%)	238 (5.1%)	106 (5.2%)	0 (0%)
2001	347 (9.2%)	128 (2.7%)	114 (5.6%)	0 (0%)
2002	362 (9.6%)	113 (2.4%)	116 (5.7%)	0 (0%)
2003	0 (0%)	354 (7.5%)	242 (12%)	0 (0%)
2004	0 (0%)	486 (10%)	115 (5.7%)	0 (0%)
2005	0 (0%)	365 (7.8%)	232 (11%)	0 (0%)
2006	247 (6.6%)	236 (5.0%)	0 (0%)	115 (19%)
2007	246 (6.6%)	0 (0%)	119 (5.9%)	232 (39%)
2008	245 (6.5%)	121 (2.6%)	239 (12%)	0 (0%)
2009	363 (9.7%)	130 (2.8%)	117 (5.8%)	0 (0%)
2010	249 (6.6%)	358 (7.6%)	0 (0%)	0 (0%)
2011	121 (3.2%)	120 (2.5%)	0 (0%)	0 (0%)
2012	240 (6.4%)	0 (0%)	0 (0%)	123 (21%)
2013	235 (6.3%)	363 (7.7%)	0 (0%)	0 (0%)
2014	243 (6.5%)	251 (5.3%)	118 (5.8%)	0 (0%)
2015	123 (3.3%)	376 (8.0%)	124 (6.1%)	0 (0%)
2016	0 (0%)	503 (11%)	0 (0%)	122 (21%)
Sex				
Female	1,769 (47%)	2,222 (47%)	948 (47%)	279 (47%)
Male	1,984 (53%)	2,484 (53%)	1,074 (53%)	313 (53%)

Table 1: Summary Statistics of Marathon Performance by Flag *(continued)*

Characteristic	**White** N = 3,753	**Green** N = 4,706	**Yellow** N = 2,022	**Red** N = 592
Age				
Mean	47.25	46.39	45.66	44.84
Min, Max	14.00, 91.00	14.00, 91.00	14.00, 90.00	14.00, 84.00
Percent off Current Record				
Mean	47.55	48.45	50.50	53.36
Min, Max	-2.25, 368.51	-1.26, 350.40	-1.42, 299.32	1.40, 247.86
Dry Bulb Temp (C)				
Mean	7.36	13.71	20.22	24.42
Min, Max	2.00, 11.00	8.74, 19.33	15.72, 28.14	22.67, 25.67
Wet Bulb Temp (C)				
Mean	3.82	9.87	15.70	19.93
Min, Max	-1.27, 9.51	5.52, 14.59	13.67, 19.02	17.54, 21.60
Relative Humidity (%)				
Mean	42.52	45.52	27.40	66.63
Min, Max	0.31, 98.25	0.28, 98.33	0.29, 87.36	50.57, 76.33
Black Globe Temp (C)				
Mean	17.34	25.71	33.21	38.84
Min, Max	9.51, 25.75	13.93, 31.92	24.96, 39.73	33.17, 44.45
Solar Radiation (W/m²)				
Mean	449.06	516.89	592.48	631.39
Min, Max	141.37, 848.12	142.73, 833.18	235.08, 909.47	344.33, 852.69
Dew Point (C)				
Mean	-1.26	6.12	12.82	17.71
Min, Max	-7.43, 8.75	-2.57, 14.23	8.43, 16.20	13.51, 20.33
Wind Speed (km/h)				
Mean	12.07	8.78	9.49	7.33
Min, Max	0.00, 21.75	3.00, 16.57	3.78, 18.20	5.33, 9.33
Wet Bulb Globe Temp (C)				
Mean	6.87	13.42	19.65	24.16
Min, Max	1.35, 9.60	10.01, 17.84	18.04, 22.79	23.20, 25.13
[1] n (%)				

Independent Variables

Figure 2a reveals that participants aged 14 to 19 and those over 75 are significantly underrepresented in the dataset, with relatively uniform participation across other age groups.

Figure 2b shows that the age distribution of age by gender was very similar between male and female participants, but it should be noted that the number of female participants in the 14-19 age group and the 80-91 age group was lower than that of male participants.

Figure 2c indicates that only a very small number (almost none) of participants managed to break the race records for their respective gender and marathon course. Most participants' finishing times were within 70% of the current records, with a minority exceeding the best times by up to 100%.

Figure 2d shows that in a few instances, races occurred under conditions where the Dew Point (DP) and Wet Bulb Temperature (Tw) dropped below 0 degrees Celsius. The Wet Bulb Globe Temperature (WBGT) averaged around 13 degrees Celsius.

Figure 2a. Distribution of Age

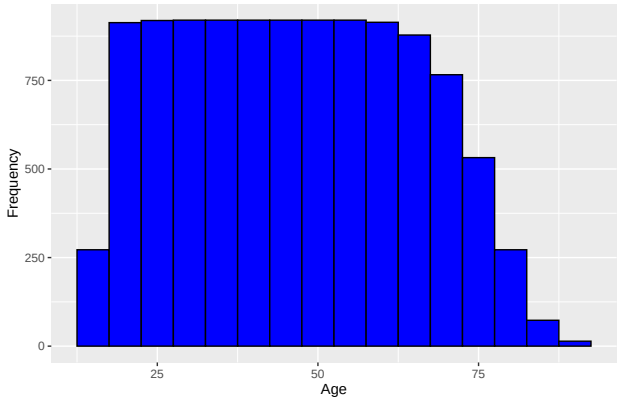


Figure 2b. Distribution of Age

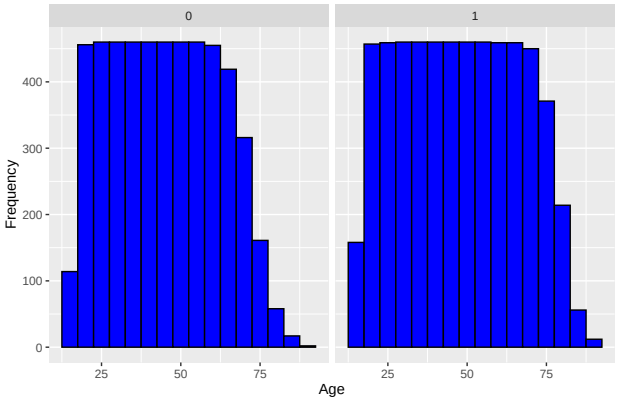


Figure 2c. Distribution of Percent off

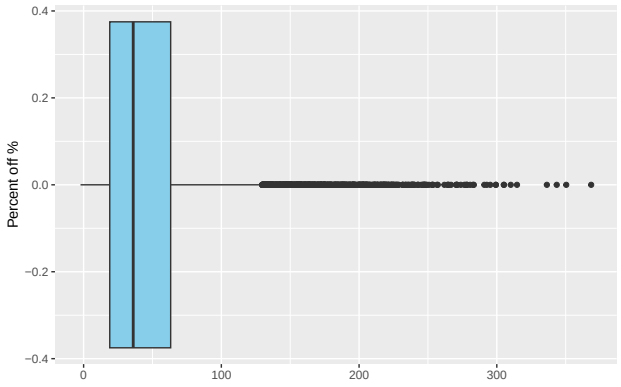
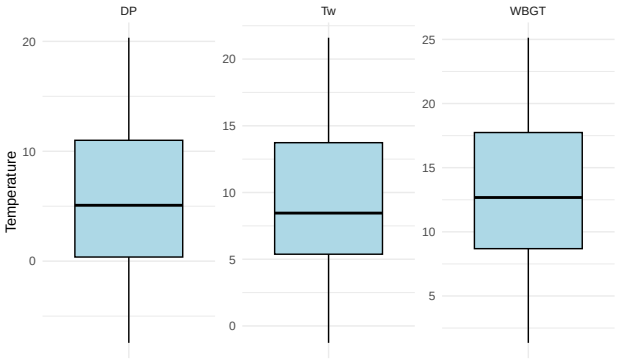


Figure 2d. Distribution of Temperature Variables



Given that both wet bulb temperature and dew point temperature are associated with relative humidity, and exhibit positive correlations with it, Figures 3a and 3b were utilized to visualize these relationships. This visualization aids in verifying the accuracy of the data recorded for Tw and DP, particularly when these temperatures are noted to be below 0 degrees Celsius.

Figures 3a and 3b demonstrate a positive correlation between Tw and relative humidity (rh), and DP and rh, respectively. These positive correlations confirm the accuracy of the data collection process.

Figure 3a. Relationship between Relative Humidity and Wet Bulb

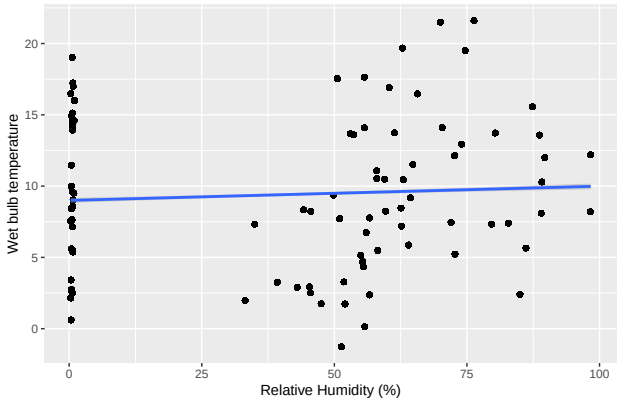
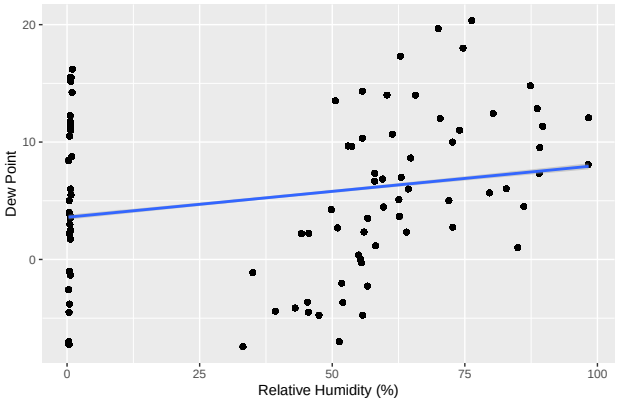


Figure 3b. Relationship between Relative Humidity and Dew Point



Similarly, Figure 4 further illustrates the correctness of data collection by showing the positive correlation between solar radiation and black globe temperature.

Figure 4. Relationship between Solar radiation and Wet Bulb

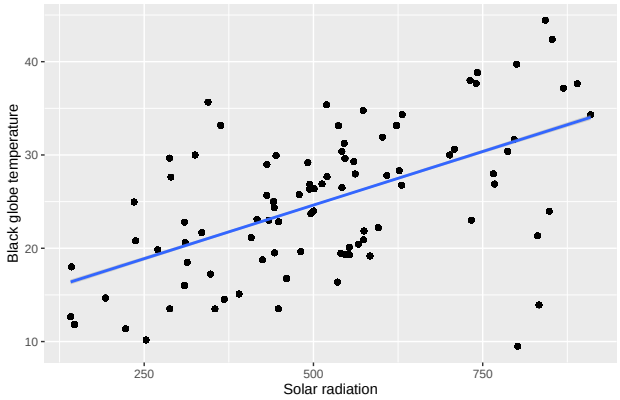


Figure 5 shows the changes in marathon performance with increasing age. We found that as age increasing, the performance deteriorates and shows a U-shaped trend. And it seems that there is no significant difference in this trend by gender.

Figure 5. Age vs. Marathon Performance by Sex

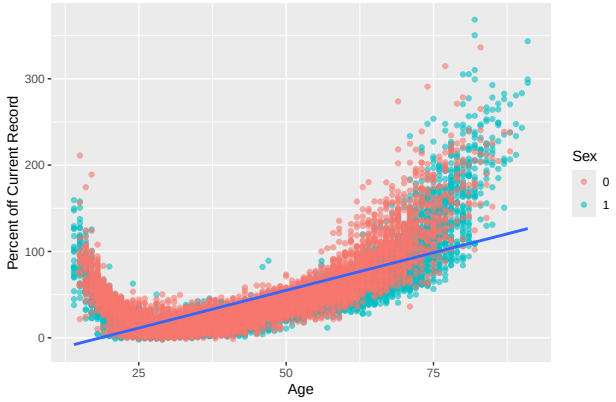


Table 2 shows the performance of each age group (mean ± standard deviation) similar to Figure 5, indicating that regardless of gender, performance improves first (best in the 20-39 age range) and then deteriorates with age. In addition, the table also shows that except for males performing worse than females in the age group of 80-91, males perform better than females in all other age groups.

Table 2: Age Group vs. Average Performance by Sex

Sex	14-19	20-29	30-39	40-49	50-59	60-69	70-79	80-91
0	62.55 ± 25.04	20.36 ± 13.64	16.5 ± 10.55	29.13 ± 11.54	50.21 ± 14.83	86.01 ± 28.92	131.3 ± 40.81	191.44 ± 46.06
1	56.38 ± 28.89	14.47 ± 12.29	13.64 ± 9.13	24.34 ± 9.36	40.01 ± 10.06	65.61 ± 19.62	117.01 ± 40.33	195.45 ± 56.24

Figure 6a presents the impact of Wet Bulb Globe Temperatur on marathon performance across different age groups. The analysis reveals that for participants aged 14 to 79 years, performance deteriorates with increasing WBGT, irrespective of gender. However, for participants aged 80 to 91 years, performance improves with rising WBGT. This counterintuitive trend might be attributed to differences in thermoregulatory mechanisms between older and younger individuals. It is also important to consider that the sample size for participants aged 80 to 91 is significantly smaller than that for other age groups, which could introduce bias into the findings.

Figure 6b examines the effect of solar radiation on performance. Similar to the findings related to WBGT, solar radiation negatively impacts the performance of participants aged 14 to 79 years. Conversely, for those aged 80 to 91 years, increased solar radiation appears to enhance performance. The trends are consistent across both genders.

Figure 6c analyzes the impact of wind speed on performance, which contrasts with the temperature-related variables. For participants between 14 to 79 years, faster wind speeds are associated with better performance, likely due to the cooling effects of higher wind speeds under equivalent conditions. For men aged 80 to 91, performance declines with higher wind speeds, whereas for women of the same age group, performance improves. This gender difference in the oldest age group suggests varying responses to wind conditions.

Figure 6d explores the relationship between relative humidity and marathon performance. For runners aged 14 to 69 years, there is a general trend where higher humidity correlates with poorer performance, although this trend is less pronounced outside the 14 to 19-year-old cohort. Remarkably, for runners aged 70 to 91 years, higher humidity levels appear to improve performance. This pattern is distinct and shows minimal gender disparity, suggesting unique adaptations or tolerances developed by older athletes towards humidity.

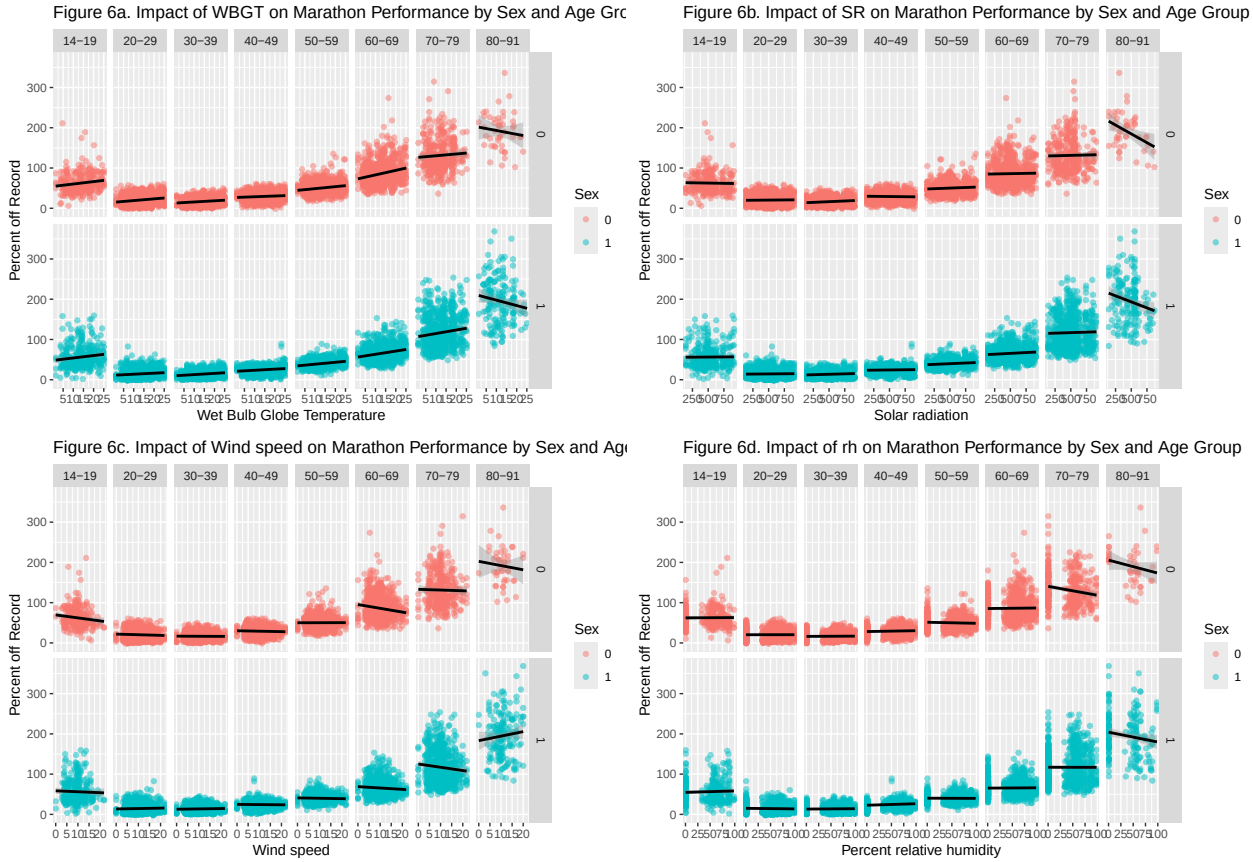


Table 3 provides a summary analysis of the impact of different WBGT levels on performance across various age groups and genders. As observed in Figure 6a, performance generally deteriorates with rising temperatures for both males and females. Notable exceptions include males aged 14-19, who perform better under red conditions than yellow, and all participants aged 80-91, whose performance under yellow and red conditions surpasses that under white and green conditions.

To further analyze the impact of environmental variables on performance and consider the effects of age and gender, we fitted a linear model. Because WBGT = (0.7 * Tw) + (0.2 * Tg) + (0.1 * Td), we will include WBGT in the model instead of separately analyzing Tw, Tg, and Td. We also know that Flag is the categorical variable of WBGT and will not consider it. Therefore, in the end, we only consider Age, Sex, rh, SR, DP, WindWBGT, as well as the two-way interaction terms of Age, Sex, and these variables. Due to the

Table 3: Performance by Flag

	14-19		20-29		30-39		40-49	
Flag	F	M	F	M	F	M	F	M
White	57.92 ± 27.28	52.41 ± 28.72	17.7 ± 12.68	13.51 ± 12.86	14.59 ± 9.51	12.05 ± 8.18	27.85 ± 10.15	23.06 ± 8.44
Green	63.05 ± 25.28	55.33 ± 27.46	20.16 ± 13.16	13.89 ± 11.4	16.64 ± 10.56	13.81 ± 8.76	29.18 ± 11.54	24.07 ± 8.08
Yellow	65.16 ± 23.03	63 ± 32.02	23.53 ± 14.62	15.93 ± 13.05	17.63 ± 11.09	14.58 ± 10.11	30.07 ± 13.44	24.68 ± 10.04
Red	69.63 ± 17.52	58.64 ± 26.75	27.71 ± 15.13	19.98 ± 11.29	23.44 ± 11.36	18.89 ± 11.44	33.48 ± 11.75	33.25 ± 15.37

	50-59		60-69		70-79		80-91	
Flag	F	M	F	M	F	M	F	M
White	47.59 ± 13.82	37.65 ± 8.75	78.48 ± 23.48	61.2 ± 15.99	129.96 ± 41.41	113.39 ± 41.52	191.15 ± 37.39	201.84 ± 59.
Green	49.99 ± 13.59	39.42 ± 9.9	86.95 ± 30.09	65.32 ± 19.89	130.22 ± 43.36	114.71 ± 37.5	195.72 ± 54.34	193.39 ± 57.
Yellow	53.21 ± 16.67	43.55 ± 10.22	94.87 ± 31.26	70.52 ± 22.48	136.5 ± 33.61	125.73 ± 42.05	186.62 ± 40.46	189.09 ± 44.
Red	58 ± 18.86	47.29 ± 11.97	98.95 ± 30.88	78.43 ± 18.73	138.18 ± 27.31	134.69 ± 42.62	140.65 ± NA	166.87 ± 30.

negative value and obvious skewness of the percent off variable, we first add 4 to the variable to adjust all values to positive, and then take the logarithm to obtain the new variable percent off log. We use percent off log to fit a weighted linear model and use the reciprocal of the square of the residuals as weights because of heteroscedasticity.

Table 4 shows that environmental conditions significantly influence marathon performance and that these effects vary based on gender and age. Specifically, high humidity, SR, Wind Speed and temperature (high WBGT) adversely affect performance, while higher dew points may be associated with better performance. Moreover, the impact of environmental conditions appears to lessen with age, and gender differences suggest that males may be more sensitive to certain environmental conditions(except DP).

Specifically, the coefficient of rh is 0.003891, suggesting that for each unit increase in relative humidity, the log-transformed marathon performance increases by approximately 0.0039 units, indicating that higher humidity correlates with slower performance. The coefficient of SR is 0.0003121, indicating that each unit increase in solar radiation increases the log-transformed performance by about 0.0003 units. The coefficient of DP is -0.005864, showing that each unit increase in dew point decreases the log-transformed performance by about 0.0059 units, suggesting that higher dew points generally correlate with faster performance. The coefficient of Wind is 0.0008482, indicating that each unit increase in wind speed raises the log-transformed performance by approximately 0.0008 units. The coefficient of WBGT is 0.03738, demonstrating that each unit increase in WBGT raises the log-transformed performance by about 0.0374 units, signifying that higher WBGT typically correlates with slower performance.

The coefficient of Age:Sex1 is 0.002694, suggesting that as age increases, the rate of performance decline is slightly higher in males compared to females. The coefficient for Sex1:rh is 0.0007550, indicating that, relative to females, males experience a greater negative impact on performance in more humid conditions. The coefficient for Age:rh is -0.00007363, meaning that the negative impact of humidity on performance diminishes with age.

Figure 7 shows that the correlation between weather variables and marathon performance is very low, indicating that a single variable of weather conditions may not have a significant individual impact on performance, or there may be a non-linear relationship between weather conditions and performance, and simple linear correlation measures cannot fully capture their relationship. Similarly, Figure 8 indicates similar marathon performance under different flag conditions.

Figure 7. Heatmap of Correlation Between Percent off and Weather C

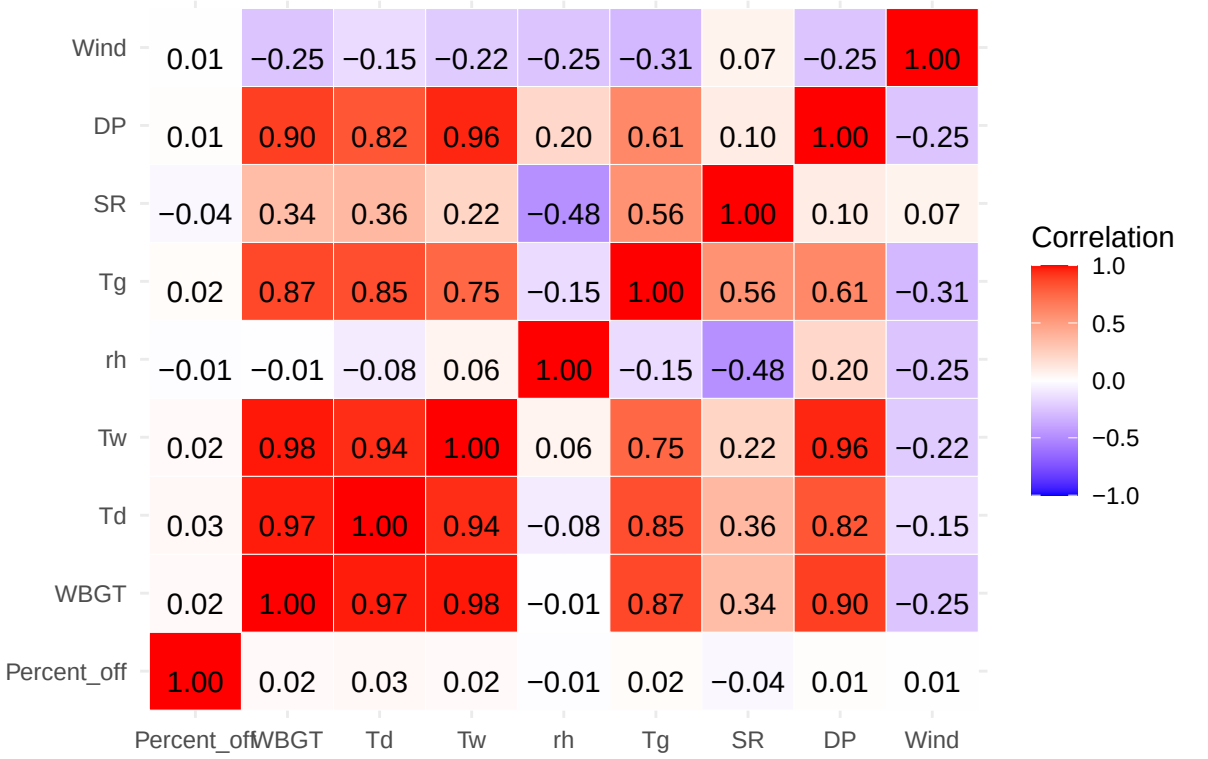
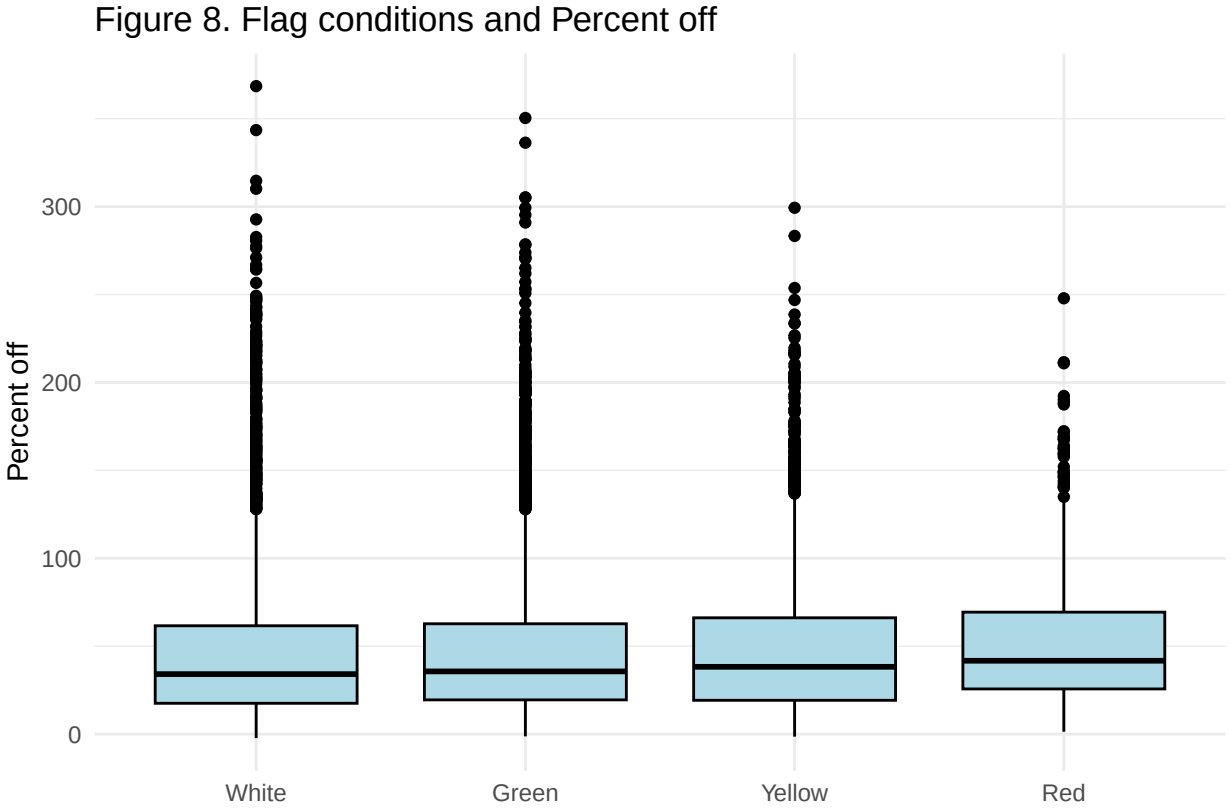


Table 4: Stepwise Regression Results: Impact of Removing Variables on Model Fit

Variable	RSS	AIC
none	11061	9.7
Tg	11228	173.6
rh	11388	330.9
Tw	11421	362.8
DP	11580	515.8
Wind	14958	3349.9
Td	16247	4265.2
Flag	23320	8263.0
SR	23609	8403.6



We utilized stepwise regression to analyze which environmental conditions have the greatest impact on marathon performance. Table 4 indicates that removing SR (Solar Radiation) resulted in an increase in the residual sum of squares (RSS) to 23,609 and an increase in the AIC to 8,403.6. This suggests that SR is one of the most influential variables contributing to the model. Additionally, Flag (indicating environmental risk level based on WBGT), Td (Dry Bulb Temperature), and Wind also have substantial impacts on marathon performance.

Further analysis was conducted by standardizing all variables to obtain coefficients and p-values. Note that WBGT and Flag were run in separate models due to their high collinearity with Tg, Tw, and Td. Table 5 shows that Td and DP (Dew Point) are the most impactful variables on marathon performance, as they have the largest coefficients and the highest t-values, indicating their strong influence within the model. SR

Table 5: Coefficient Analysis without WBGT and Flag

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.6368	0.0002	19000.4347	0
Tw	0.0289	0.0036	8.1304	0
Tg	0.0123	0.0007	18.2245	0
Td	0.0961	0.0016	58.7926	0
rh	0.0115	0.0003	35.2491	0
SR	-0.0484	0.0003	-153.2738	0
DP	-0.0752	0.0022	-34.3974	0
Wind	0.0164	0.0002	98.7630	0

Table 6: Coefficient Analysis with WBGT

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.5680	0.0007	4942.8136	0
FlagGreen	0.0763	0.0009	86.6834	0
FlagYellow	0.1217	0.0018	69.0909	0
FlagRed	0.2566	0.0023	111.9388	0
rh	0.0071	0.0002	32.7153	0
SR	-0.0544	0.0003	-209.2608	0
DP	-0.0991	0.0004	-262.1034	0
Wind	0.0308	0.0003	120.0723	0
WBGT	0.1025	0.0010	106.8107	0

follows closely, with both its t-value and coefficient underscoring its importance.

Table 6 highlights that FlagRed and WBGT are the most significant variables affecting performance. Both exhibit large coefficients and high t-values, demonstrating that extreme temperature conditions and high WBGT levels are associated with significantly impaired marathon performance. SR (Solar Radiation) and DP (Dew Point) show significant negative effects in the model, suggesting that under certain environmental conditions, they may help improve performance, potentially by optimizing the heat dissipation mechanisms of the athletes.

Discussion

In this study, we analyzed the effects of age, gender, and environmental conditions on marathon performance and identified the weather parameters most significantly impacting performance. Despite yielding valuable insights, the study has the following three limitations:

1. Data Scope and Sample Representativeness The dataset includes data from a limited number of major marathons (e.g., Boston, New York, Chicago), which may not be representative of all marathon events globally. Factors such as geographical location, altitude, and seasonal variations could influence environmental conditions, thereby limiting the generalizability of these findings. Moreover, the dataset only includes the best single-age performances within a specified age range (14 to 91 years), which may not fully reflect the overall distribution of performances across age groups. Therefore, the results might primarily apply to the most elite athletes within this sample.

2. Measurement of Environmental Variables and Multicollinearity The environmental variables used in the study (e.g., WBGT, dry bulb temperature, wet bulb temperature, and black globe temperature) exhibit high multicollinearity, making it challenging to disentangle the independent effects of each

variable accurately. Although stepwise regression and variable standardization were applied to mitigate the impact of multicollinearity, this strong correlation may still affect the stability of the coefficients for some variables. Consequently, the true effects of certain environmental factors may be masked or misrepresented in terms of their magnitude and direction.

3. Lack of Consideration for Individual Variability and Potential Confounders The study primarily focuses on the overall impact of environmental variables on marathon performance but does not account for individual-level differences such as training status, heat acclimatization ability, body composition, hydration strategies, or pacing tactics, all of which could significantly influence performance outcomes. Additionally, the dataset lacks detailed time-series information on dynamic environmental variables like wind speed and solar radiation, preventing an evaluation of their real-time effects during races. The absence of these individual and dynamic factors may limit a deeper understanding of marathon performance under various environmental 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(gtsummary)
library(RColorBrewer)
library(naniar)
library(dplyr)
library(knitr)
library(tidyr)
library(reshape2)
library(kableExtra)
# read in data
data <- readRDS('/Users/fusei/Desktop/24FALL/PHP2550/Project1/project1.cleaned_data.rds')

### MISSING DATA
# remove subjects with no data from current study
## missing pattern
gg_miss_fct(data,Race) + labs(title = "Figure 1a. NA in Variables and Race")
gg_miss_fct(data,Year) + labs(title = "Figure 1b. NA in Variables and Year")

data <- data[complete.cases(data),]
#dim(data) ## 11073 x 14

data_table1 <- data %>%
  mutate(
    Race = factor(Race, levels = 0:4, labels = c("Boston", "Chicago", "NYC", "Twin Cities", "Grandma's"
    Year = factor(Year),
    Sex = factor(Sex, levels = c(0, 1), labels = c("Female", "Male")),
    Flag = factor(Flag, levels = c("White", "Green", "Yellow", "Red"))
  )


table1 <- tbl_summary(
  data_table1,
  by = "Flag",
  type = list(
    Age ~ "continuous2",
    c(Td, Tw, rh, Tg, SR, DP, Wind, WBGT) ~ "continuous2",
    Percent_off ~ "continuous2"
  ),
  statistic = list(
    all_continuous() ~ c("{mean}", "{min}, {max}")
  ),
  digits = list(
    all_continuous() ~ c(2, 2)
  ),
  label = list(
    Race ~ "Race",
    Sex ~ "Sex",
    Year ~ "Year",
    Age ~ "Age",
    Percent_off ~ "Percent off Current Record",
```

```
    Td ~ "Dry Bulb Temp (C)",
    Tw ~ "Wet Bulb Temp (C)",
    rh ~ "Relative Humidity (%)",
    Tg ~ "Black Globe Temp (C)",
    SR ~ "Solar Radiation (W/m²)",
    DP ~ "Dew Point (C)",
    Wind ~ "Wind Speed (km/h)",
    WBGT ~ "Wet Bulb Globe Temp (C)"
  )
)  %>%
# Table 1

   as_kable_extra(booktabs = TRUE,
                  caption = "Summary Statistics of Marathon Performance by Flag",
                 longtable = TRUE, linesep = "") %>%
  kableExtra::kable_styling(font_size = 8,
                            latex_options = c("repeat_header", "HOLD_position"))

table1
## single variable
## age
ggplot(data, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  ggtitle("Figure 2a. Distribution of Age") +
  xlab("Age") +
  ylab("Frequency")
##age by gender
ggplot(data, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  ggtitle("Figure 2b. Distribution of Age") +
  facet_wrap(~Sex) +
  xlab("Age") +
  ylab("Frequency")

## performance
ggplot(data, aes(x = Percent_off)) +
  geom_boxplot(fill = "skyblue") +
  ggtitle("Figure 2c. Distribution of Percent off") +
  xlab("") +
  ylab("Percent off %")

## tw dp + WBGT
p <- ggplot(data, aes(x = value)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +
  theme_minimal() +
  labs(x = "Temperature ", y = "Frequency") +
  facet_wrap(~variable, scales = "free")

data_long <- data %>%
  select(Tw, DP, WBGT) %>%
  pivot_longer(cols = everything(), names_to = "variable", values_to = "value")

ggplot(data_long, aes(x = "", y = value)) +
```

```r
  geom_boxplot(fill = "lightblue", color = "black") +
  facet_wrap(~variable, scales = "free_y") +
  labs(title = "Figure 2d. Distribution of Temperature Variables",
       x = "",
       y = "Temperature ") +
  theme_minimal()

# dp and rh
ggplot(data, aes(x = rh, y = Tw)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Figure 3a. Relationship between Relative Humidity and Wet Bulb", x = "Relative Humidity
# tw and rh
ggplot(data, aes(x = rh, y = DP)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Figure 3b. Relationship between Relative Humidity and Dew Point", x = "Relative Humidit


ggplot(data, aes(x = SR, y = Tg)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(title = "Figure 4. Relationship between Solar radiation and Wet Bulb", x = "Solar radiation", y =


# Age vs. Performance by sex
ggplot(data, aes(x = Age, y = Percent_off)) +
  geom_point(aes(color = Sex), alpha = 0.6) +
  geom_smooth(method = "lm") +
  labs(title = "Figure 5. Age vs. Marathon Performance by Sex", x = "Age", y = "Percent off Current Rec

data_table2 <- data %>%
  mutate(Age_Group = case_when(
    Age >= 14 & Age <= 19 ~ "14-19",
    Age >= 20 & Age <= 29 ~ "20-29",
    Age >= 30 & Age <= 39 ~ "30-39",
    Age >= 40 & Age <= 49 ~ "40-49",
    Age >= 50 & Age <= 59 ~ "50-59",
    Age >= 60 & Age <= 69 ~ "60-69",
    Age >= 70 & Age <= 79 ~ "70-79",
    Age >= 80 & Age <= 91 ~ "80-91",
    TRUE ~ "Other"
  ))
summary_table <- data_table2 %>%
  group_by(Sex, Age_Group) %>%
  summarise(
    Mean_Performance = mean(Percent_off, na.rm = TRUE),
    SD_Performance = sd(Percent_off, na.rm = TRUE),
    .groups = 'drop'
  ) %>%
  mutate(Performance = paste(round(Mean_Performance, 2), "±", round(SD_Performance, 2)))

summary_table <- summary_table[,c(1,2,5)]
```

```r
summary_wide <- summary_table %>%
  pivot_wider(names_from = Age_Group, values_from = Performance)

# kable(summary_table, caption = "Table 2. Age Group vs. Average Performance by Sex",
#       col.names = c("Sex", "Age Group", "Average Performance ± SD"),
#       align = 'c', format = "html") %>%
#  kableExtra::kable_styling(bootstrap_options = c("striped", "hover"))
kable_output <- summary_wide %>%
  kable("latex", booktabs = TRUE, longtable = TRUE, caption = "Age Group vs. Average Performance by Sex
    kableExtra::kable_styling(font_size = 8, latex_options = c("repeat_header", "HOLD_position"))
kable_output
data_figure6 <- data %>%
  mutate(Age_Group = case_when(
    Age >= 14 & Age <= 19 ~ "14-19",
    Age >= 20 & Age <= 29 ~ "20-29",
    Age >= 30 & Age <= 39 ~ "30-39",
    Age >= 40 & Age <= 49 ~ "40-49",
    Age >= 50 & Age <= 59 ~ "50-59",
    Age >= 60 & Age <= 69 ~ "60-69",
    Age >= 70 & Age <= 79 ~ "70-79",
    Age >= 80 & Age <= 91 ~ "80-91",
    TRUE ~ "Other"
  ))
### Temperature Variables (Td, Tw, WBGT) vs. Performance:
ggplot(data_figure6, aes(x = WBGT, y = Percent_off)) +
  geom_point(aes(color = Sex), alpha = 0.5) +
  geom_smooth(method = "lm", color = "black") +
  facet_grid(Sex ~ Age_Group) +
  labs(title = "Figure 6a. Impact of WBGT on Marathon Performance by Sex and Age Group",
       x = "Wet Bulb Globe Temperature", y = "Percent off Record")

ggplot(data_figure6, aes(x = SR, y = Percent_off)) +
  geom_point(aes(color = Sex), alpha = 0.5) +
  geom_smooth(method = "lm", color = "black") +
  facet_grid(Sex ~ Age_Group) +
  labs(title = "Figure 6b. Impact of SR on Marathon Performance by Sex and Age Group",
       x = "Solar radiation", y = "Percent off Record")

ggplot(data_figure6, aes(x = Wind, y = Percent_off)) +
  geom_point(aes(color = Sex), alpha = 0.5) +
  geom_smooth(method = "lm", color = "black") +
  facet_grid(Sex ~ Age_Group) +
  labs(title = "Figure 6c. Impact of Wind speed on Marathon Performance by Sex and Age Group",
       x = "Wind speed", y = "Percent off Record")

### Relative Humidity (rh) vs. Performance:

ggplot(data_figure6, aes(x = rh, y = Percent_off)) +
  geom_point(aes(color = Sex), alpha = 0.5) +
  geom_smooth(method = "lm", col = "black") +
  facet_grid(Sex ~ Age_Group) +
  labs(title = "Figure 6d. Impact of rh on Marathon Performance by Sex and Age Group",
       x = "Percent relative humidity", y = "Percent off Record")
```

```r
# Figure 4
data_table3 <- data %>%
  mutate(Age_Group = case_when(
    Age >= 14 & Age <= 19 ~ "14-19",
    Age >= 20 & Age <= 29 ~ "20-29",
    Age >= 30 & Age <= 39 ~ "30-39",
    Age >= 40 & Age <= 49 ~ "40-49",
    Age >= 50 & Age <= 59 ~ "50-59",
    Age >= 60 & Age <= 69 ~ "60-69",
    Age >= 70 & Age <= 79 ~ "70-79",
    Age >= 80 & Age <= 91 ~ "80-91",
    TRUE ~ "Other"
  )) %>%
  group_by(Flag, Sex, Age_Group) %>%
  summarise(
    Mean_Percent_off = mean(Percent_off, na.rm = TRUE),
    SD_Percent_off = sd(Percent_off, na.rm = TRUE),
    .groups = 'drop'
  ) %>%
  mutate(Percent_off_Summary = paste(round(Mean_Percent_off, 2), "±", round(SD_Percent_off, 2)))

data_table3 <- data_table3 %>%
  mutate(Age_Sex = str_c("Age_", Age_Group, "_Sex_", Sex))
data_wide <- data_table3 %>%
  pivot_wider(names_from = Age_Sex, values_from = Percent_off_Summary)

data_table3 <- data.frame(Flag = unique(data_wide$Flag))

for(var in c("Age_14-19_Sex_0"
,"Age_20-29_Sex_0" , "Age_30-39_Sex_0" , "Age_40-49_Sex_0",  "Age_50-59_Sex_0" , "Age_60-69_Sex_0" , "Ag
    data_table3[[var]] <- data_wide[!is.na(data_wide[[var]]),][[var]]
}

data_table3 <- data_table3[,c("Flag","Age_14-19_Sex_0","Age_14-19_Sex_1","Age_20-29_Sex_0","Age_20-29_S
                              "Age_30-39_Sex_0" , "Age_30-39_Sex_1", "Age_40-49_Sex_0", "Age_40-49_Sex_
data_table3_1 <-data_table3[,1:9]
data_table3_2 <-data_table3[,c(1,10:17)]
kable_output3_1 <- kable(data_table3_1, format = "latex", booktabs = TRUE,
                      align = rep('c', 9),
                      col.names = rep("", 9),
                      caption = "Performance by Flag") %>%
  kable_styling(latex_options = c("striped", "landscape"), full_width = F, font_size = 8) %>%
  add_header_above(header = c("Flag" = 1, "F" = 1, "M" = 1, "F" = 1, "M" = 1, "F" = 1, "M" = 1,
                              "F" = 1, "M" = 1 )) %>%
  add_header_above(header = c(" " = 1, "14-19" = 2, "20-29" = 2, "30-39" = 2, "40-49" = 2
                              ))

kable_output3_2 <- kable(data_table3_2, format = "latex", booktabs = TRUE,
                      align = rep('c', 9),
                      col.names = rep("", 9)) %>%
  kable_styling(latex_options = c("striped", "landscape"), full_width = F, font_size = 8) %>%
  add_header_above(header = c("Flag" = 1, "F" = 1, "M" = 1, "F" = 1, "M" = 1, "F" = 1, "M" = 1,
```

```
                              "F" = 1, "M" = 1)) %>%
  add_header_above(header = c(" " = 1,
                              "50-59" = 2, "60-69" = 2, "70-79" = 2, "80-91" = 2))

kable_output3_1
kable_output3_2

data$Percent_off_log = log(data$Percent_off + 4)
model <- lm(Percent_off_log ~ Age * Sex +
               rh + SR + DP + Wind + WBGT +
               Sex * rh + Sex * SR + Sex * DP + Sex * Wind + Sex * WBGT +
               Age * rh + Age * SR + Age * DP + Age * Wind + Age * WBGT,
            data = data)
weights <- 1 / residuals(model)^2
wls_model <- lm(Percent_off_log ~ Age * Sex +
                    rh + SR + DP + Wind + WBGT +
                    Sex * rh + Sex * SR + Sex * DP + Sex * Wind + Sex * WBGT +
                    Age * rh + Age * SR + Age * DP + Age * Wind + Age * WBGT,
                data = data, weights = weights)
coefficients_df <- as.data.frame(coef(summary(wls_model)))
kable_table <- kable(coefficients_df, format = "latex", caption = "Regression Results Summary",
                     col.names = c("Estimate", "Std. Error", "t value", "Pr(>|t|)"),
                     digits = 4)
kable_styled <- kable_table %>%
    kable_styling(bootstrap_options = c("striped", "hover"), full_width = F) %>%
    column_spec(1, bold = T, border_right = T) %>%
    column_spec(4, color = "red")
kable_styled
## continuous variables
cor_matrix <- cor(data[, c("Percent_off", "WBGT", "Td", "Tw","rh"
                                    ,"Tg","SR","DP","Wind")], use = "complete.obs")
melted_cor_matrix <- melt(cor_matrix)
ggplot(melted_cor_matrix, aes(Var1, Var2, fill = value)) +
    geom_tile(color = "white") +
    geom_text(aes(label = sprintf("%.2f", value)), vjust = 1) +
    scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1, 1), spa
    theme_minimal() +
    labs(x = "", y = "", title = "Figure 7. Heatmap of Correlation Between Percent off and Weather Condi

## categorical variable
ggplot(data, aes(x = Flag, y = Percent_off)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Figure 8. Flag conditions and Percent off",
       x = "",
       y = "Percent off") +
  theme_minimal()

data$Percent_off_log = log(data$Percent_off + 4)
model <- lm(Percent_off_log ~ Tw + Tg + Td + Flag +
               rh + SR + DP + Wind + WBGT ,
            data = data)
weights <- 1 / residuals(model)^2
wls_model <- lm(Percent_off_log ~ Tw + Tg + Td + Flag +
```

```
                    rh + SR + DP + Wind + WBGT ,
                  data = data, weights = weights)

backward_model <- step(wls_model, direction = "backward",trace = 0)

## step output
stepwise_results <- data.frame(
  Variable = c("none","Tg", "rh", "Tw", "DP", "Wind", "Td", "Flag", "SR"),
  RSS = c(11061,11228, 11388, 11421, 11580, 14958, 16247, 23320, 23609),
  AIC = c(9.7, 173.6, 330.9, 362.8, 515.8, 3349.9, 4265.2, 8263.0, 8403.6)
)


kable_table <- kable(
  stepwise_results,
  format = "latex",
  col.names = c("Variable", "RSS", "AIC"),
  caption = "Stepwise Regression Results: Impact of Removing Variables on Model Fit"
) %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = F, position = "center", font_si

kable_table
data$Percent_off_log = log(data$Percent_off + 4)
data_scaled <- data.frame(scale(data[, c("Tg", "rh", "Tw", "DP", "Wind", "Td", "SR", "WBGT")]))
data_scaled$Percent_off_log <- data$Percent_off_log
data_scaled$Flag <- data$Flag
## no WBGT
model <- lm(data_scaled$Percent_off_log ~ Tw + Tg + Td +
                rh + SR + DP + Wind  ,
            data = data_scaled)
weights <- 1 / residuals(model)^2
wls_model <- lm(data_scaled$Percent_off_log ~ Tw + Tg + Td +
                    rh + SR + DP + Wind  ,
                  data = data_scaled, weights = weights)
coefficients_df <- as.data.frame(coef(summary(wls_model)))
kable_table <- kable(coefficients_df, format = "latex", caption = "Coefficient Analysis without WBGT an
                     col.names = c("Estimate", "Std. Error", "t value", "Pr(>|t|)"),
                     digits = 4)
kable_styled <- kable_table %>%
    kable_styling(bootstrap_options = c("striped", "hover"), full_width = F) %>%
    column_spec(1, bold = T, border_right = T) %>%
    column_spec(4, color = "red")
kable_styled
## WBGT
modelw <- lm(data_scaled$Percent_off_log ~  Flag +
                rh + SR + DP + Wind + WBGT ,
            data = data_scaled)
weights <- 1 / residuals(modelw)^2
wls_modelw <- lm(data_scaled$Percent_off_log ~  Flag +
                    rh + SR + DP + Wind + WBGT ,
                  data = data_scaled, weights = weights)

coefficients_df <- as.data.frame(coef(summary(wls_modelw)))
```

```
kable_table <- kable(coefficients_df, format = "latex", caption = "Coefficient Analysis with WBGT ",
                     col.names = c("Estimate", "Std. Error", "t value", "Pr(>|t|)"),
                     digits = 4)
kable_styled <- kable_table %>%
    kable_styling(bootstrap_options = c("striped", "hover"), full_width = F) %>%
    column_spec(1, bold = T, border_right = T) %>%
    column_spec(4, color = "red")
kable_styled
```