IMPACT OF ENVIRONMENTAL FACTORS ON MARATHON PERFORMANCE ACROSS AGE AND GENDER

PHP2550 Project 1: An exploratory data analysis

Yanwei (Iris) Tong

2024-09-29

Abstract

Purpose: The objectives of this exploratory study were to determine the effects of gender, age, and environmental stressors on marathon performance, and to identify the factors with maximal impacts.

Methods: We analyzed the results of 96 major marathon events in the U.S. from 1993 to 2016. LOESS plots, Spearman correlation test, GAM with a Gamma family, LM with log transformed outcomes, and other supplementary analyses were perfromed to visualize and investigate the effects of age, weather conditions– i.e., Wet Bulb Globe Temperatur (°C), wind speed (km/hr), solar radiation (W/m^2), percent relative humidity (%)– and the concentrations of 4 criteria air pollutants: $PM_{2.5}$ ($\mu g/m^3$), SO_2 (ppm), NO_2 (ppm), and O_3 (ppb) on marathon performance for all runners and record holders across genders.

Results and conclusion: Marathon performance exhibited a U-shaped relationship with age, where middle-aged runners (20-40 years old) performed best. This pattern was significantly modified by gender and moderately affected by WBGT Flag level. Environmental factors showed a relatively consistent impact between genders, but varied by age group. General runners aged 20–60 were relatively unaffected by environmental conditions, while younger and older runners showed greater sensitivity. Although several environmental factors were statistically associated with performance, no matter measured as % off course records and net race time for general runners or course records for top performers, the variability explained by the environmental factors were small, indicating limited practical impact. This suggests that environmental conditions, while statistically significant, may play a minimal role in determining overall marathon performance.

INTRODUCTION

The relationship between environmental factors and marathon performance has been the subject of extensive research. Ely et al. (2007) has shown that increasing Wet Bulb Globe Temperature (WBGT) negatively affects overall speed. Interestingly, the impact of weather appears less pronounced for female runners, as Vihma (2010) noted, suggesting potential physiological differences. Deaner et al. (2015) further emphasized that men are more likely than women to slow over the course of a marathon, possibly due to pacing strategies or differences in heat tolerance. Age also plays a critical role, with older runners exhibiting less variance in marathon pace compared to younger participants (Nikolaidis and Knechtle (2019)). However, environmental stressors like higher temperatures and humidity affect both older male and female runners, as shown in the New York City Marathon, where they experienced significant performance declines (Knechtle et al. (2021)). Beyond heat, air temperature and pollution also factor into performance outcomes. El Helou et al. (2012) demonstrated that the relationship between air temperature and marathon performance follows a quadratic trend, with optimal running temperatures varying by performance level. As temperatures exceed these optimal ranges, performance declines and withdrawal rates increase, with Ozone (O_3) levels potentially compounding these effects.

Together, these studies underscore the complex interplay of gender, age, and environmental conditions on marathon performance, highlighting the need for a comprehensive examination of how these factors influence endurance running. Therefore, in this project, by examining data from 96 race events of 5 major marathon races in the U.S. from 1993 to 2016, we would like to re-explore the effects and patterns of the following factors on marathon performance: 1) gender 2) the increase in age, and 3) environmental parameters, including weather conditions and air quality, and eventually to identify the parameters with the maximal influences.

METHODS

Data Overview

Marathon race results were obtained from five major U.S. marathons (Boston, Chicago, New York City, Grandmas, and Twin Cities) spanning the years 1993 to 2016. The total number of performances collected was 11,564. The gender distribution was nearly even across all, with 47% female and 53% male participants across all events. The average age of runners varied slightly between marathons, ranging from 44 to 50 years, with the Twin Cities marathon hosting a slightly younger average age of participants compared to the others. Statistically significant differences in age were observed amongst races ($p < 0.001$).

Table 1: Participant Characteristics by Marathon Race, N = 11564

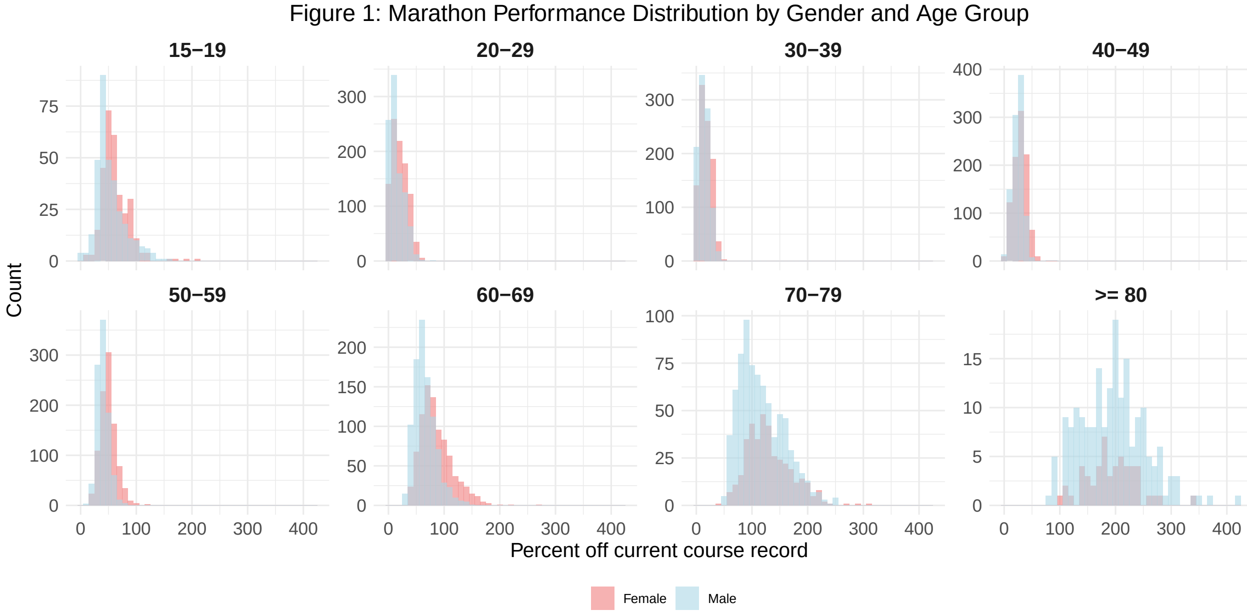
Characteristic	**Boston** N = 2,088	**Chicago** N = 2,930	**Grandmas** N = 2,000	**NYC** N = 2,553	**Twin Cities** N = 1,993	**p-value**
Gender						0.8
Female	984 (47%)	1,402 (48%)	934 (47%)	1,210 (47%)	922 (46%)	
Male	1,104 (53%)	1,528 (52%)	1,066 (53%)	1,343 (53%)	1,071 (54%)	
Age	46.96 (17.26)	49.57 (18.76)	44.03 (17.51)	45.82 (17.89)	44.72 (17.44)	<0.001

[1] n (%); Mean (SD)

[2] Pearson's Chi-squared test; Kruskal-Wallis rank sum test

The distributions of performance, measured as % off current course record by gender, as illustrated in **Figure 1**, were right-skewed across all age groups, with most runners finishing closer to the course record, and a long tail of slower times, resembling the characteristics of Gamma distributions. This skewness is consistent across both genders, though differences in the central tendency of performance between men and women were observed within each age group. Given the non-normal distribution of net race times, transformation and adaptation will be necessary (e.g., using log transform or Gamma family) in subsequent analyses to ensure more accurate modeling and interpretation of the factors influencing marathon performance.

Figure 1: Marathon Performance Distribution by Gender and Age Group



Relevant weather conditions of the 96 event dates were also provided. These are 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/hr), and Wet Bulb Globe Temperature (WBGT,°C). WBGT was also categorized to 5 flag levels, quantifying risk of heat illness: White = WBGT <10C, Green = WBGT 10-18 °C, Yellow = WBGT >18-23 °C, Red = WBGT >23-28 °C, and Black = WBGT >28 °C. The air pollution concentrations of sulfur dioxide (parts per million), nitrogen dioxide (parts per million), ozone (parts per billion), and particular matter 2.5 (μg/m^3) were extracted from the EPA meta database using the `R` package `RAQSAPI`. SO_2, NO_2, and O_3 were derived from daily averages of 1-hour measurements, whereas $PM_{2.5}$ was based on a 24-hour monitoring.

Regarding the distribution of the environmental parameters, the small p-values across **Table 2** indicate significant differences in weather conditions and air quality among the five major marathons. Variables such as WBGT, wind speed, humidity, and pollutants like NO_2 and SO_2 demonstrate considerable variability across locations, highlighting the unique environmental contexts in which each race was conducted. These findings suggest that environmental conditions may play a role in influencing marathon performance, particularly across different regions.

Table 2: Summary of Environmental Factors by Marathon, N = 96

Characteristic	**Boston** N = 18	**Chicago** N = 23	**Grandmas** N = 17	**NYC** N = 21	**Twin Cities** N = 17	**p-value**
WBGT (°C)						<0.001
Mean (SD)	11.3 (4.6)	10.7 (5.0)	18.6 (3.3)	12.1 (5.9)	13.3 (5.6)	
Min, Max	6.5, 23.2	3.7, 18.9	14.0, 25.1	1.3, 24.7	6.5, 24.2	
WBGT flag						
White	9 (50%)	11 (50%)	0 (0%)	6 (30%)	5 (31%)	
Green	7 (39%)	7 (32%)	6 (38%)	12 (60%)	7 (44%)	
Yellow	1 (5.6%)	4 (18%)	8 (50%)	1 (5.0%)	3 (19%)	
Red	1 (5.6%)	0 (0%)	2 (13%)	1 (5.0%)	1 (6.3%)	
Black	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	
Td (°C)						<0.001
Mean (SD)	11.6 (6.0)	11.7 (4.8)	18.9 (3.4)	12.4 (6.2)	13.2 (5.7)	
Min, Max	5.3, 28.1	5.3, 20.0	13.0, 24.6	2.0, 25.7	7.0, 24.7	
Tw (°C)						<0.001
Mean (SD)	7.6 (3.9)	7.6 (5.1)	14.9 (2.5)	8.6 (5.9)	9.9 (5.6)	
Min, Max	2.5, 17.5	0.6, 17.0	10.0, 19.7	-1.3, 21.5	2.0, 21.6	
Tg (°C)						0.003
Mean (SD)	24.2 (8.5)	21.4 (6.1)	31.6 (8.1)	24.5 (6.5)	25.0 (6.8)	
Min, Max	9.5, 42.4	11.4, 34.8	13.9, 44.5	10.2, 35.7	12.7, 35.4	
DP (°C)						<0.001
Mean (SD)	3.3 (4.5)	2.7 (7.2)	12.4 (3.3)	4.7 (7.1)	6.0 (7.5)	
Min, Max	-4.4, 13.5	-7.3, 16.2	4.0, 18.0	-7.0, 19.7	-7.4, 20.3	
Wind speed (Km/hr)						0.010
Mean (SD)	12.0 (4.6)	11.2 (4.7)	9.2 (2.9)	8.2 (3.3)	8.8 (3.3)	
Min, Max	4.8, 21.8	0.0, 20.0	3.8, 14.0	3.0, 16.3	3.7, 15.7	
% relative humidity						0.008
Mean (SD)	34.9 (35.2)	26.8 (31.2)	48.9 (35.5)	60.6 (10.7)	41.4 (35.4)	
Min, Max	0.3, 98.3	0.3, 98.3	0.4, 89.7	43.0, 85.0	0.4, 89.1	
SR (W/m^2)						<0.001
Mean (SD)	654.0 (191.3)	401.2 (134.0)	679.3 (195.3)	459.7 (96.3)	436.5 (142.9)	
Min, Max	147.2, 852.7	142.7, 573.4	289.5, 909.5	252.8, 608.5	141.4, 630.2	
NO2 (ppb)						<0.001
Mean (SD)	9.2 (3.3)	18.6 (5.6)	0.9 (1.3)	16.4 (7.6)	7.0 (2.8)	
Min, Max	4.0, 14.7	8.6, 28.4	0.0, 2.5	3.8, 28.9	3.7, 13.7	
SO2 (ppb)						<0.001
Mean (SD)	2.3 (1.4)	3.5 (2.0)	1.1 (1.4)	3.7 (2.8)	0.7 (0.7)	
Min, Max	0.3, 4.9	1.1, 7.9	0.0, 3.1	0.1, 10.4	0.0, 2.9	
O3 (ppm)						<0.001
Mean (SD)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	
Min, Max	0.0, 0.1	0.0, 0.0	0.0, 0.1	0.0, 0.0	0.0, 0.0	
PM2.5 (microgram/m^3)						0.015
Mean (SD)	9.2 (5.2)	12.8 (7.6)	6.1 (2.5)	9.3 (7.3)	6.5 (5.2)	

Table 2: Summary of Environmental Factors by Marathon, N = 96 *(continued)*

Characteristic	**Boston** N = 18	**Chicago** N = 23	**Grandmas** N = 17	**NYC** N = 21	**Twin Cities** N = 17	**p-value**
Min, Max	2.2, 21.0	4.4, 30.9	2.6, 8.8	2.3, 31.3	2.4, 22.3	

[1] n (%)
[2] Kruskal-Wallis rank sum test

Figure 2 demonstrates the correlation between the environmental parameters. SO_2 and NO_2 exhibit a strong positive correlation (0.82), while ozone shows mild negative correlations with both SO_2 (-0.24) and NO_2 (-0.31). $PM_{2.5}$ correlates positively with SO_2 (0.51) and NO_2 (0.67), but less with ozone (0.12). These patterns may suggest different underlying pollution sources. Importantly, WBGT, which is a weighted average of dry bulb, wet bulb, and globe temperatures, expectedly shows strong correlations with DP (0.87), Td (0.97), Tw (0.98), and Tg (0.86). This high degree of correlation among temperature-related variables highlights the issue of multi-collinearity. As WBGT integrates these components and represents overall thermal stress, further analyses will focus on WBGT as the primary temperature measure, thereby avoiding redundancy and potential co-linearity issues in the statistical models.

Figure 2: Correlation between Environmental Parameters

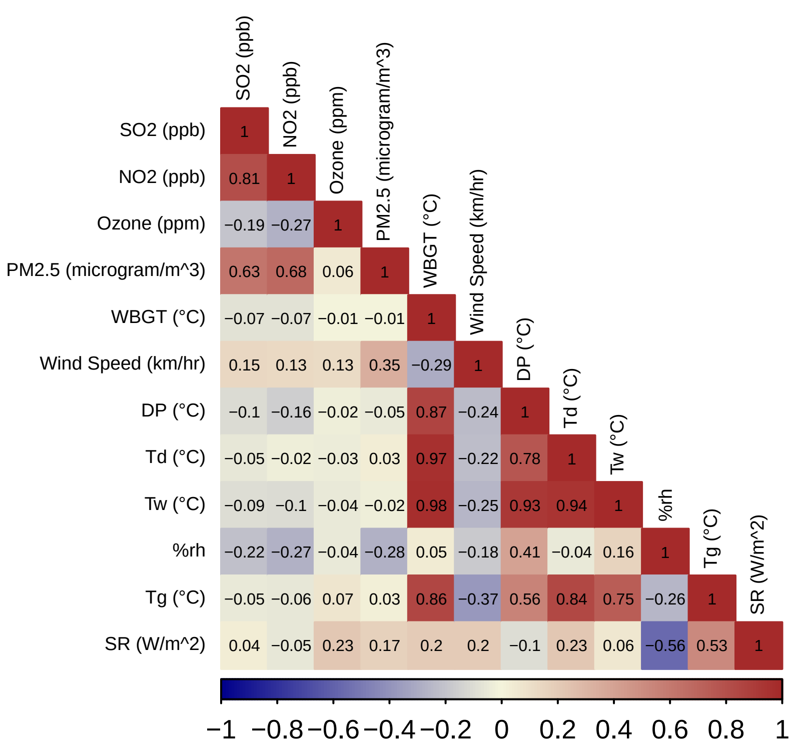
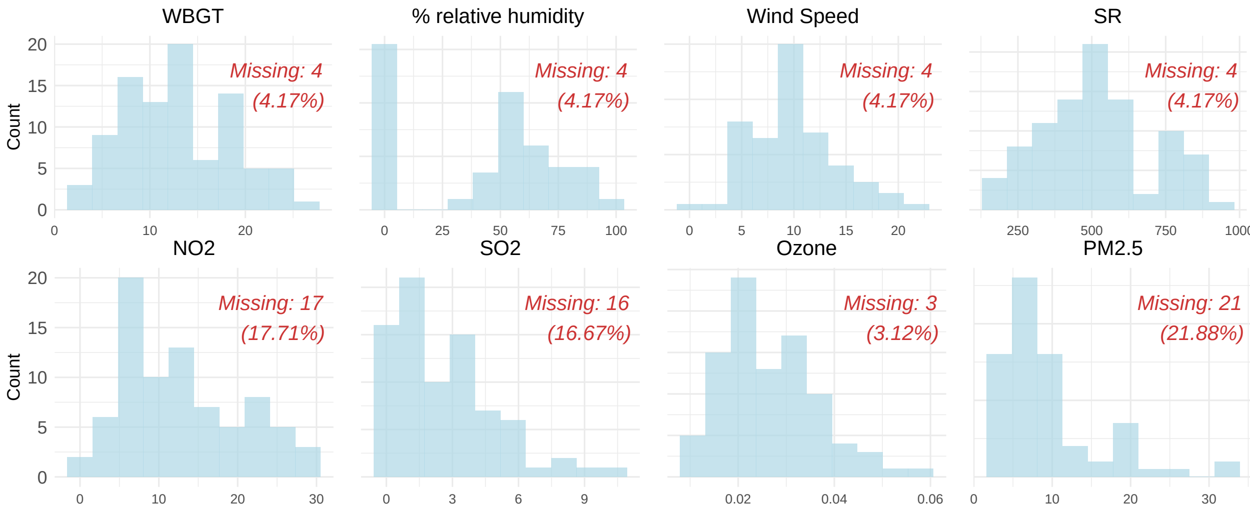


Figure 3 presents histograms of key environmental factors from the 96 races with levels of missingness. It shows that these variables are generally non-normal and with high variability. Moreover, most of the missingness in the dataset pertains to the air pollutant parameters, with NO_2 showing a missingness rate of 17.71%, SO_2 at 16.67%, and $PM_{2.5}$ at 21.88%. These higher levels of missing data may present challenges to the robustness of subsequent analyses, particularly when assessing the impact of air quality on marathon performance. In contrast, the missingness rates for weather conditions are relatively low, approximately 4%, and only 3% for O_3, both of which satisfy the 5% rule of thumb, allowing for the safe assumption that ignoring these missing data will not significantly affect the overall results.

Figure 3: Distribution of Environmental Factors with Missingness



Analysis Methods

In the main exploratory analysis, we first employed locally estimated scatterplot smoothing (LOESS) to visually investigate the potential impact of age on marathon performance between genders and effects of various environmental factors across different age-gender subgroups. These visualizations provided an initial overview of the relationships, highlighting any non-linear trends or differences in performance between genders or across age groups.

A Spearman correlation test was then performed to quantitatively assess the overall associations between environmental variables and marathon performance, both for the entire cohort and for record-setting runners. Spearman correlation is suitable for capturing monotonic relationships without assuming normality, given that we knew non-linearity and non-normality exist for most of the parameters and performance outcomes.

Finally, to identify the most significant predictors and provide a more robust analysis, generalized additive models (GAMs) with a Gamma family and a log link function were fitted to investigate the association between environmental factors and marathon performance in general runners. Based on the LOESS figures, linear models with log transformed course record as the outcome were fitted to understand the environmental effects on top performers. These approaches allowed us to generate approximate F-values, p-values, and adjusted R^2 for each environmental factor and enabled us to identify which factors exhibited the most significant influence on marathon performance, offering a more comprehensive understanding of the key environmental variables impacting both general and top-level performance.

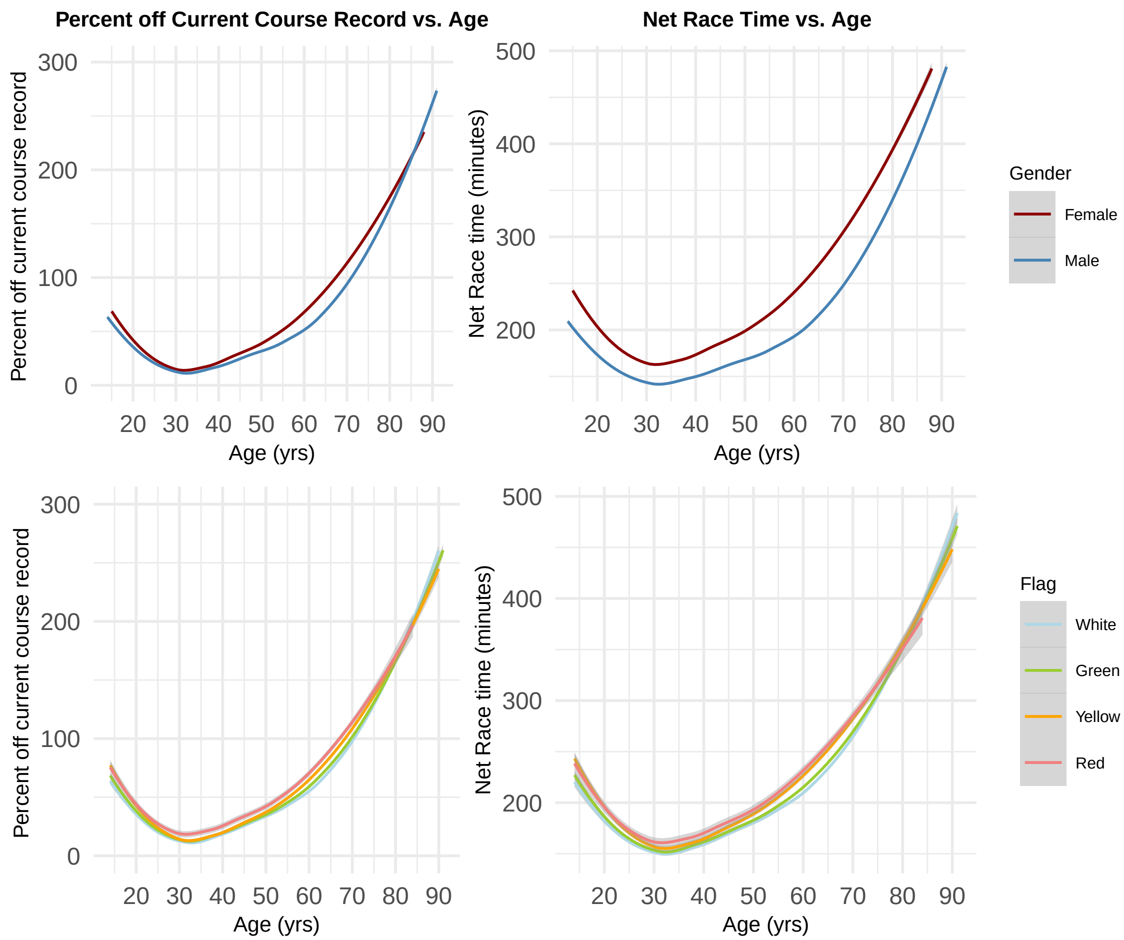
RESULTS

The Impact of Age on Marathon Performance across Genders and WBGT Flags

Figure 4 illustrates the relationship between age and marathon performance stratified by gender and WBGT Flag level, measured in two different ways: percent off the current course record and net race time (minute). Consistent with findings from preliminary analyses, we observed a U-shaped relationship between age and performance. It implies that marathon performance improves with age during the early years, peaks around the 30s, and then declines thereafter. Notably, there is a clear distinction between male and female performance, with females generally performing slower than males. The pattern remains consistent regardless of how performance is measured. However, since the percent off course record metric already adjusts for gender, the performance gap between genders is narrower in that plot compared to the net race time plot.

The performance-age relationship is also modified by the WBGT Flag, which is an indicator of heat risk. As the WBGT Flag level increases from white (low risk) to red (higher risk), a decline in performance is observed across all age groups. Specifically, runners at higher WBGT levels tend to have worse performances, as evidenced by both higher percent off the course record and longer net race times. This trend highlights the potential detrimental impact of heat stress on marathon performance, but the effect modification of WBGT Flag was not as noticeable as that of gender.

Figure 4: Impact of Age on Marathon Performance by Gender and WBGT Flag



The General Correlation between Environmental Fators and Marathon Performance for Both General and Top Runners

To explore the impact of environmental conditions on marathon performance, we started with analyzing the overall correlations between various environmental factors and these two performance metrics: net race time in all runners (n = 11564) and the course records in minutes for all race events for both genders (n = 192). The results are displayed in **Table 3**. Overall, the correlations are relatively low. Among the environmental factors, WBGT showed the highest positive correlation with both net race time and course record, indicating that higher WBGT is associated with slower performances. Conversely, wind speed, solar radiation, and % relative humidity displayed negative correlations, suggesting that higher values of these factors may slightly improve performance, although the effect sizes are small. Interestingly, while NO_2 showed a very weak positive correlation with net race time in all runners (r = 0.0022), it had a stronger negative correlation with the course record (r = -0.118). Similar opposite directional patterns were observed for $PM_{2.5}$ and solar radiation, suggesting that the association between environmental factors and marathon performance may vary depending on whether we consider the entire field of runners or focus specifically on top performers (course record setters). This differential effect highlights the need for tailored analyses when assessing environmental impacts on performance across different levels of competitiveness.

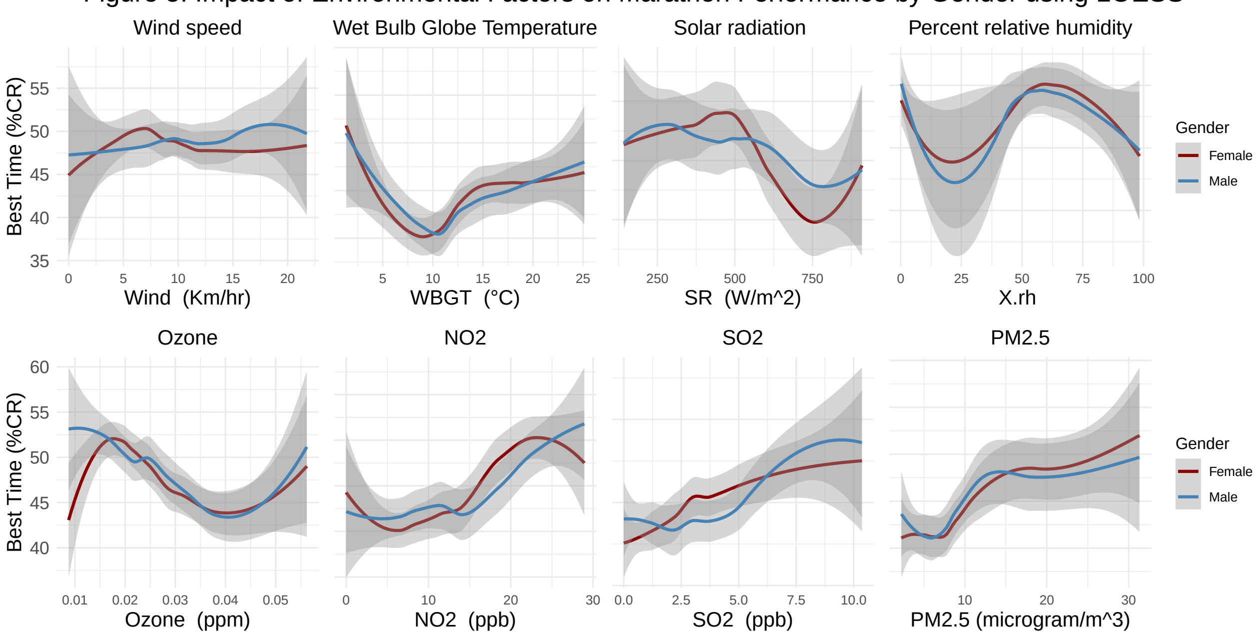
Table 3: Spearman Correlation between Environmental Factors and Net Race Time for General Runners and Record Setters

	Net race time in all runners	Course record
WBGT	0.0660	0.1274
%rh	-0.0121	-0.0630
Wind Speed	-0.0096	-0.0351
Solar Radiation	-0.0091	0.0365
NO2	0.0022	-0.1179
SO2	-0.0098	-0.1256
Ozone	-0.0311	-0.0045
PM2.5	0.0168	-0.0469

The Impact of Environmental Fators on Marathon Performance for General Runners

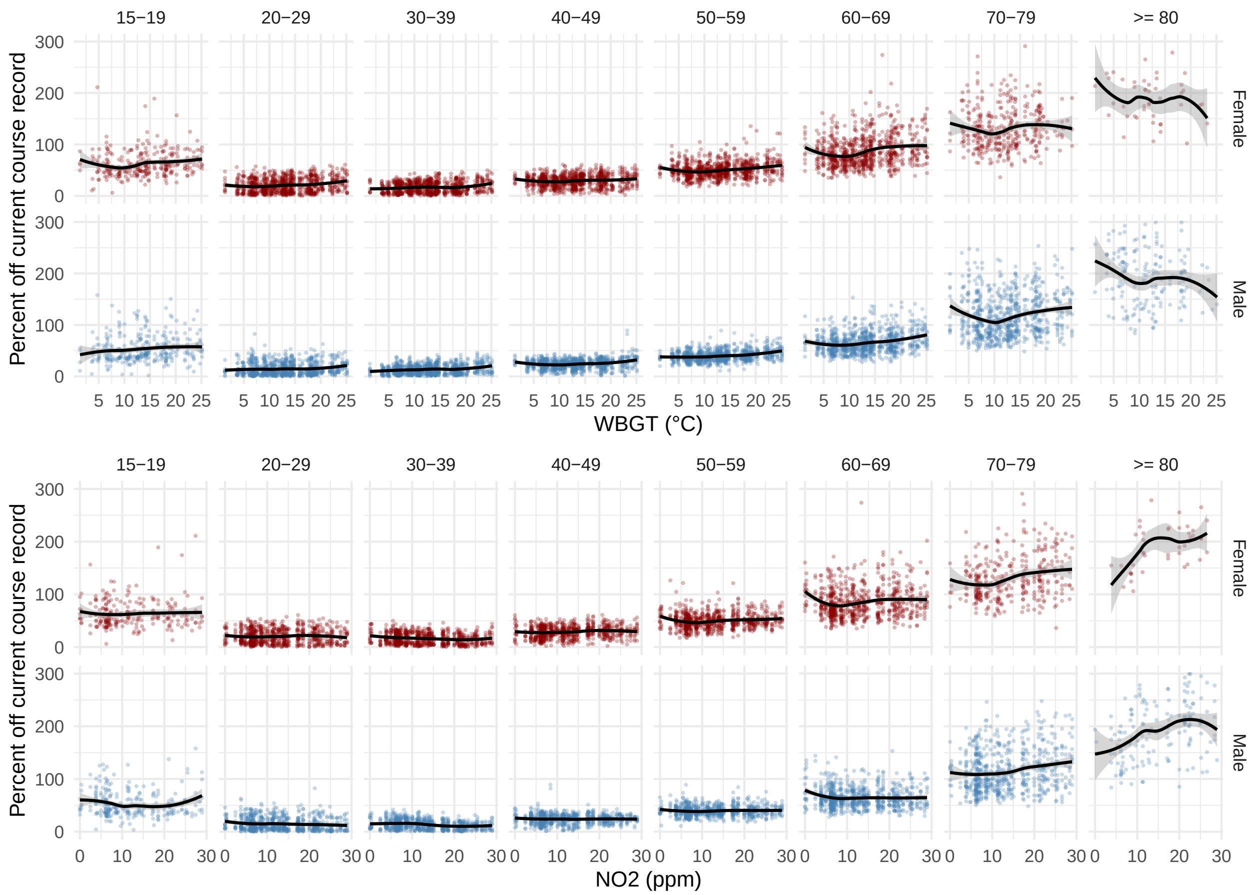
Figure 5 illustrates the impact of various environmental factors on marathon performance for all runners, stratified by gender. Across all environmental factors–including wind speed, WBGT, solar radiation, percent relative humidity, and air pollutants (Ozone, NO_2, SO_2, and $PM_{2.5}$)–the effect of gender appears minimal, with male and female runners exhibiting overlapping patterns. The impact of environmental conditions on performance is notably non-linear, with clear curvatures visible across several variables, particularly WBGT and percent relative humidity. These results suggest complex interactions between environmental factors and performance, but without a significant difference in gender-specific responses.

Figure 5: Impact of Environmental Factors on Marathon Performance by Gender using LOESS



Then, we further broke down the environmental impacts by both genders and age groups. Take the effects of WBGT and NO_2 as examples. For runners in the 20-59 year-old age groups, the two factors' influence on marathon performance appears to be minimal, as the regression line is nearly horizontal. In contrast, for both younger runners (15-19 years) and older runners (above 60 years), the relationship between WBGT or NO_2 and marathon performance seems more non-linear, with slight increases and decreases in performance with the increasing WBGT or increasing concentration of NO_2. Despite these age-specific trends, the overall patterns for men and women are similar, reiterating that gender does not significantly modify the relationship between environmental stressors and marathon performance.

Figure 6: the Impact of WBGT and NO2 on Marathon Performance for General Runners



Note that due to data sparsity, runners with age <= 14 were not included in the LOESS plots. And, for clarity, the by-age-and-gender LOESS plots for the remaining six environmental factors can be found in the **Figure Appendix** section.

Given the non-linearity of the data, generalized additive models (GAM) with a Gaussian family and log link was conducted to identify the most significant environmental factors. Note that to better visualize the results, the factors were analyzed on a standardized scale. The results presented in **Figure 7** and **Table 4** present the GAM results for general runners. Except for wind spped, the approximate p-values for other smooth terms of the GAM are very small, indicating strong statistical evidence of associations between these environmental factors and marathon performance (as measured by percent off the course record).

However, the adjusted R^2 values and the percentage of deviance explained are consistently low, with adjusted R^2 values ranging from 0.13% for percent relative humidity to a maximum of 0.78% for NO_2. Similarly, the percentage of deviance explained does not exceed 1% for any of the environmental factors, with most falling below 0.5%. In short, despite the statistical significance, the practical impacts (effect sizes) of these findings are minimal, meaning that their ability to explain variation in marathon performance is extremely limited.

Figure 7: GAM Fits of Environmental Factors on % off Course Record for General Runners

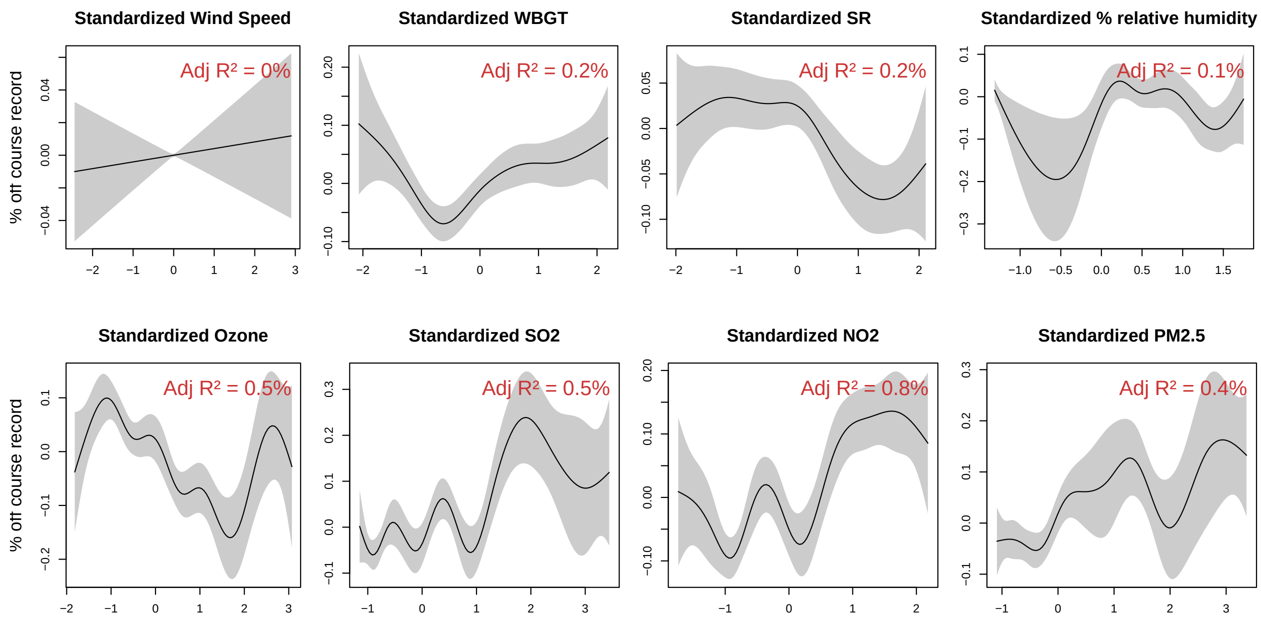


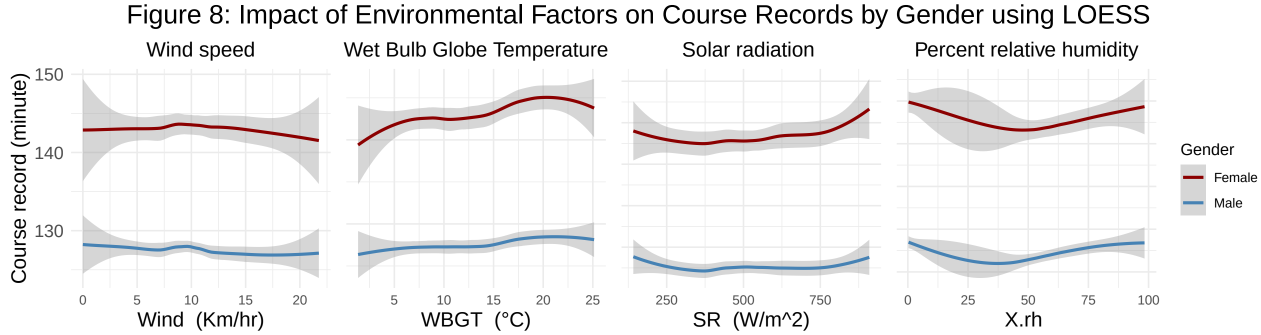
Table 4: GAM Results- Percent off Course Record for General Runners vs. Standardized Environmental Factors

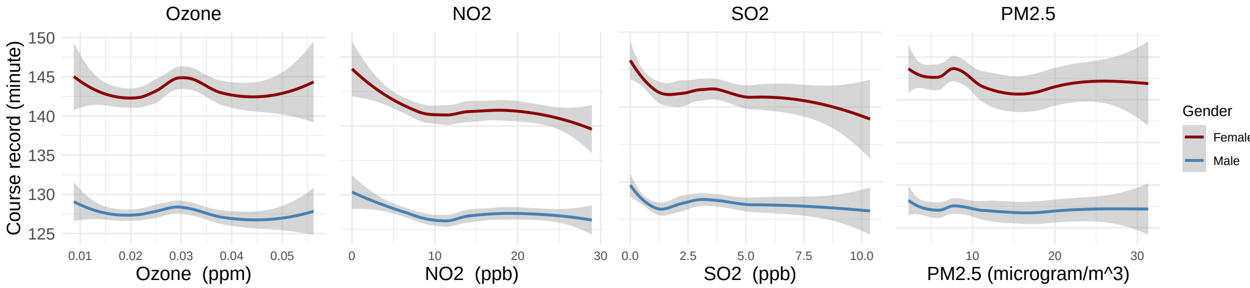
Standardized variable	n	Approxi. F-value	Approxi. p-value	Adj. R2	% Dev. explained
Wind Speed	11039	0.2193	0.6423	-0.01%	0%
WBGT	11039	4.3495	0.0002	0.22%	0.26%
SR	11039	4.1508	0.0009	0.17%	0.21%
% relative humidity	11039	2.6112	0.0095	0.13%	0.19%
Ozone	11150	7.3712	0.0000	0.48%	0.54%
SO2	9614	5.3652	0.0000	0.46%	0.51%
NO2	9508	9.4523	0.0000	0.78%	0.82%
PM2.5	9080	4.0081	0.0000	0.36%	0.41%

The Impact of Environmental Fators on Course Records

To investigate the environmental effects on course records, we performed LOESS smoothing again. **Figure 8** reveals that the impact of environmental factors on course records tends to approach linearity, especially in comparison to the more complex and clear curvatures seen with net race times for general runners. Thus, for simplicity, this subsection employed a linear model controlled for gender to assess the environmental effects.

Figure 8: Impact of Environmental Factors on Course Records by Gender using LOESS





Notably, the gender variable alone explains nearly 88% of the variability in course records, leaving limited room for other factors to contribute meaningfully. While some environmental factors, such as solar radiation, SO_2, and O_3, are statistically significant ($p < 0.05$), the approximate isolated R^2 values (the difference between full-model R^2 and gender-only-model R^2) for these factors remain negligible, indicating that the practical influence of these environmental factors on course records is minimal despite their statistical significance. Interestingly, the betas for air pollutants, including NO_2, SO2, and PM2.5, are negative, suggesting that higher air pollution levels are associated with better course records. This "counterintuitive" result is likely not a genuine relationship, as the R^2 values are below 1%, indicating that this "negative" association may be spurious or driven by confounding factors.

Table 5: Linear Model Results- log(Course Records) vs. Environmental Factors (Controlled for Gender)

Standardized Var.	Beta	CI Lower	CI Upper	p-value	Approxi. Isolated R^2
WBGT	-0.0016	-0.0047	0.0015	0.3198	0.08%
%rh	0.0053	0.0023	0.0083	0.0007	0.76%
Wind Speed	0.0020	-0.0011	0.0051	0.2035	0.12%
Solar Radiation	-0.0042	-0.0072	-0.0011	0.0071	0.48%
NO2	-0.0003	-0.0034	0.0029	0.8670	-0.41%
SO2	-0.0048	-0.0080	-0.0015	0.0042	0.34%
Ozone	-0.0051	-0.0083	-0.0019	0.0021	0.75%
PM2.5	-0.0020	-0.0055	0.0015	0.2591	-0.63%

CONCLUSION

This exploratory study investigated the relationships between age, gender, and environmental factors on marathon performance. Regarding the realtionship between performance and increase in age, a clear U-shaped trend was observed, where runners in their 20s and 30s outperformed both younger and older groups. Gender and WBGT Flag level modified this pattern, with men consistently showing better performance across all ages and higher heat risk negatively influencing performance. Environmental factors such as WBGT and air pollutants NO_2 and O_3 demonstrated age-specific effects, with minimal impact on runners aged 20 to 60, while younger and older athletes exhibited more variability in sensitivity to these conditions. However, the environmental effect patterns didn't differ notably across genders.

Although several environmental factors showed statistically significant associations with marathon performance for both general runners or course record setters, their practical impact was limited. The adjusted R-squared and deviance explained by these factors were consistently small, suggesting minimal practical influence on overall performance. Counterintuitively, air pollutants like NO_2 and SO_2 showed negative associations with course records, likely due to small effect sizes and low explained variance rather than true causal relationships. Additionally, the level of missingness in pollutant data, particularly for NO_2, SO_2, and $PM_{2.5}$, introduced challenges, highlighting the need for more robust methods like multiple imputation to address these gaps in future studies.

In conclusion, while environmental factors may have some impact on marathon performance, their effects are marginal compared to the dominant influences of age and gender. Future research should incorporate more advanced methods to address missing data and better quantify the role of environmental factors, particularly in age and gender-specific analyses.

LIMITATIONS

This analysis encountered notable levels of missingness in the air pollutant data, particularly for NO_2, SO_2, and $PM_2.5$, where the missing rate far exceeded the 5% safety line to ignore missingness. While this level of missingness does not immediately invalidate the analysis, it does present a challenge to the robustness of the conclusions. For simplicity, this study excluded entries with missing data rather than employing imputation techniques. Ideally, methods like Multiple Imputation by Chained Equations (MICE) could be used to address the missingness more rigorously, thereby preserving more data for analysis. Alternatively, fuller or more comprehensive pollutant data could be extracted from the EPA's extensive meta database, offering another avenue for future improvements.

Given the non-linear and non-normal distribution of marathon performance in relation to environmental factors, several statistical approaches were employed, including LOESS smoothing, Spearman correlation, and Generalized Additive Models. These methods allowed for flexible (non-parametric) modeling of complex, non-linear relationships. However, the use of GAM presented limitations, particularly in the ability to adjust for covariates such as age and gender. Unlike (generalized) linear models , which easily incorporate such covariates, GAM's focus on non-parametric smooth terms makes it harder to control for these factors explicitly. As a result, this analysis was unable to clearly quantify the differences in environmental effects across different age groups and genders, which limits our ability to discern how these factors might differentially affect various runner populations. Future work should explore hybrid models or extensions of methods like GAM or Spline that can better handle covariates like age and gender to provide a more complete picture.

Data Privacy and Code Availability

Primary data were provided by Dr. Brett Romano Ely and Dr. Matthew Ely from the Department of Health Sciences at Providence College. The original data cannot be shared directly for privacy. Replication scripts are available at https://github.com/YanweiTong-Iris/PHP2550-Fall24/tree/main/Project1.

Reference

Deaner, R. O., Carter, R. E., Joyner, M. J., and Hunter, S. K. (2015), "Men are more likely than women to slow in the marathon," *Medicine and science in sports and exercise*, NIH Public Access, 47, 607.

El Helou, N., Tafflet, M., Berthelot, G., Tolaini, J., Marc, A., Guillaume, M., Hausswirth, C., and Toussaint, J.-F. (2012), "Impact of environmental parameters on marathon running performance," *PloS one*, Public Library of Science San Francisco, USA, 7, e37407.

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

Knechtle, B., McGrath, C., Goncerz, O., Villiger, E., Nikolaidis, P. T., Marcin, T., and Sousa, C. V. (2021), "The role of environmental conditions on master marathon running performance in 1,280,557 finishers the 'new york city marathon'from 1970 to 2019," *Frontiers in Physiology*, Frontiers Media SA, 12, 665761.

Nikolaidis, P. T., and Knechtle, B. (2019), "Do fast older runners pace differently from fast younger runners in the 'new york city marathon'?" *The Journal of Strength & Conditioning Research*, LWW, 33, 3423–3430.

Vihma, T. (2010), "Effects of weather on the performance of marathon runners," *International journal of biometeorology*, Springer, 54, 297–306.

Figure Appendix

Supplemental LOESS plots for the impact of environmental factors on marathon performance by age and gender

Figure S1: The Impact of Solar Radiation on Marathon Performance by Age and Gender

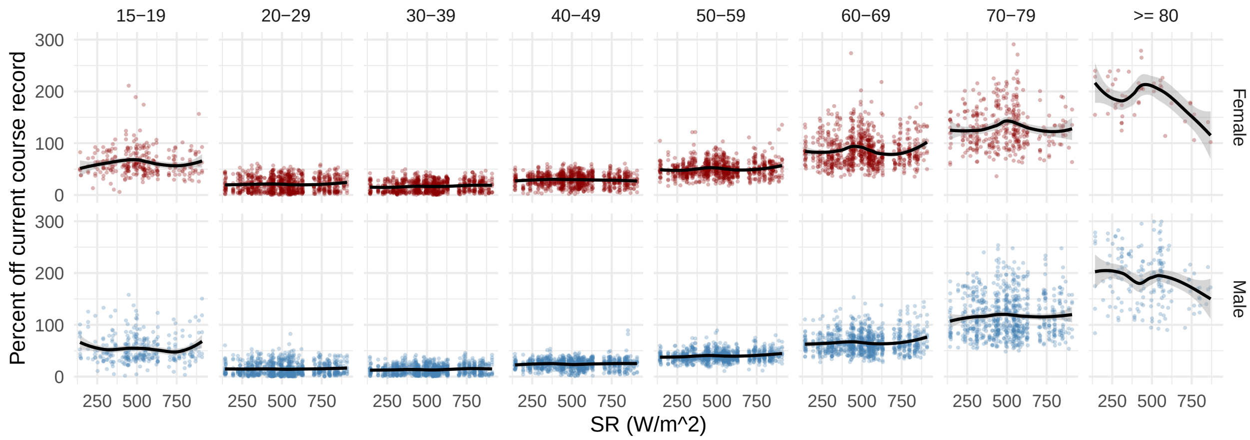


Figure S2: The Impact of Percent Relative Humidity on Marathon Performance by Age and Gender

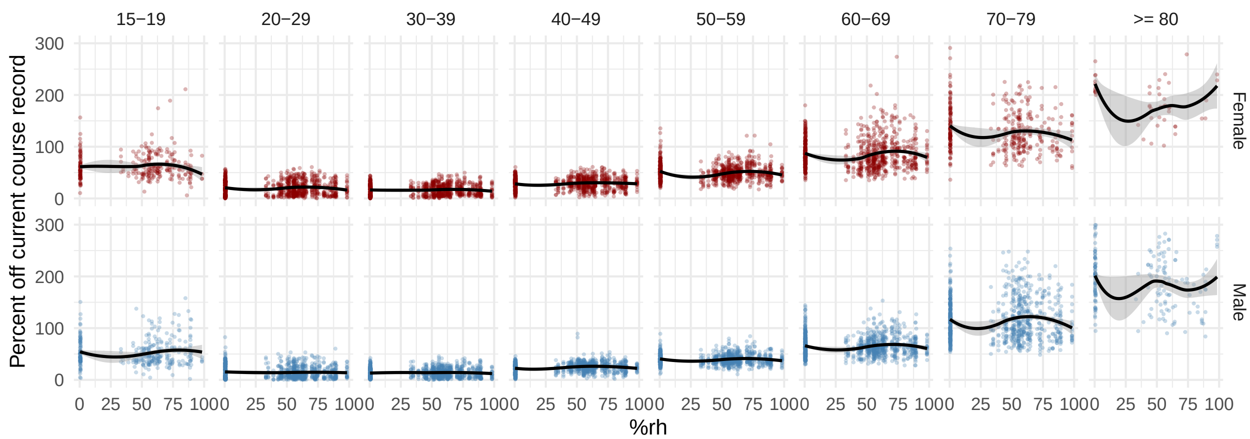


Figure S3: The Impact of Wind Speed on Marathon Performance by Age and Gender

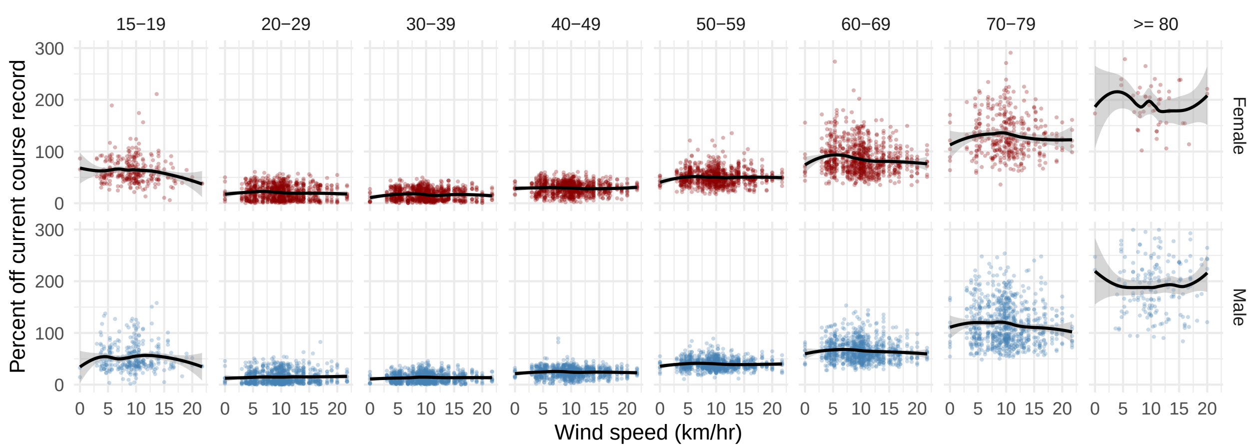


Figure S4: The Impact of SO2 on Marathon Performance by Age and Gender

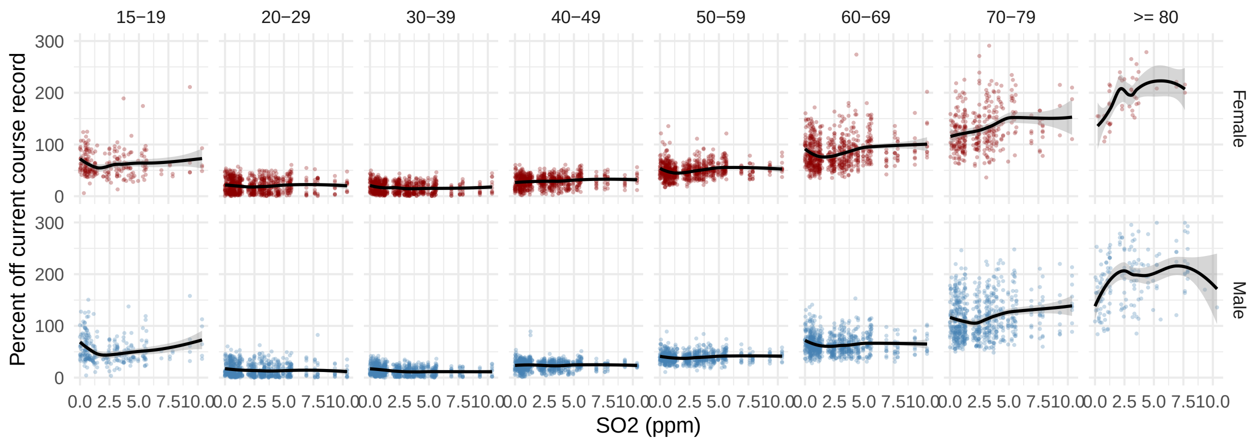


Figure S5: The Impact of Ozone on Marathon Performance by Age and Gender

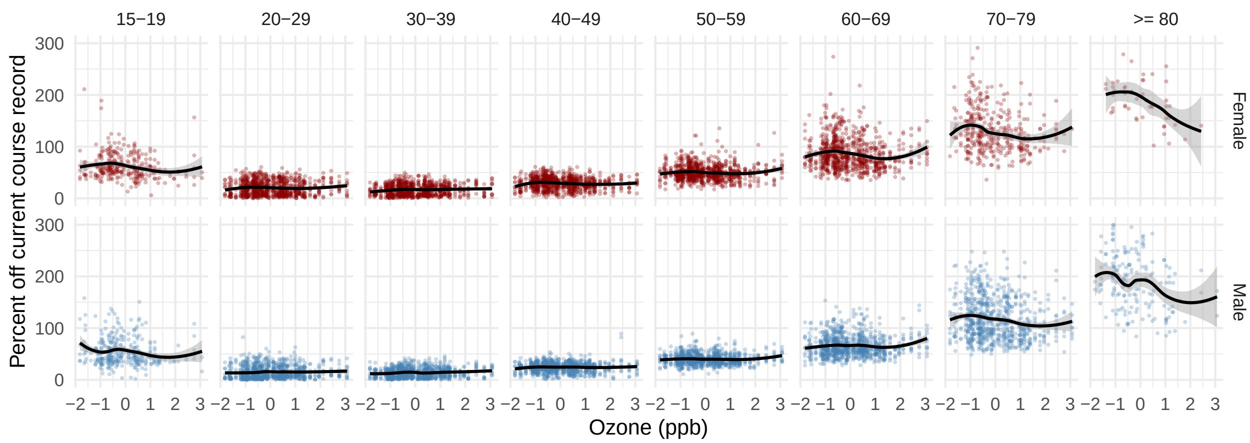
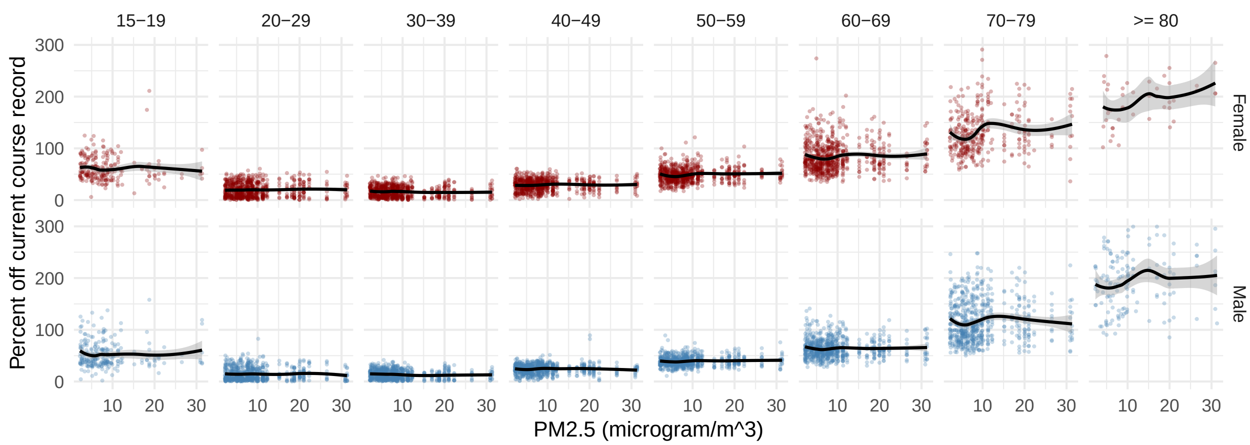


Figure S6: The Impact of PM2.5 on Marathon Performance by Age and Gender



Code Appendix

```
# Set up knit environment
knitr::opts_chunk$set(echo = FALSE,
                      message = FALSE,
                      warning = FALSE,
                      error = FALSE)

# Load necessary packages
library(tidyverse)
library(kableExtra)
library(knitr)
library(ggplot2)
library(naniar)
library(gtsummary)
library(gt)
library(patchwork)
library(stargazer)
library(knitcitations)
library(mosaic)
library(summarytools)
library(npreg)
library(mgcv)


# Define data path
data_path = "/Users/yanweitong/Documents/PHP2550-Data/Project1"

# Import datasets
main_data = read.csv(paste0(data_path, "/project1.csv"))
aqi_data = read.csv(paste0(data_path, "/aqi_values.csv"))
record_data = read.csv(paste0(data_path, "/course_record.csv"))
# Merge main and record data sets
record_data <- record_data %>%
  mutate(Sex = ifelse(Gender == "F", 0, 1)) %>%
  mutate(Race_code = case_when(Race == "B"~0,
                               Race == "C"~1,
                               Race == "NY"~2,
                               Race == "TC"~3,
                               Race == "D" ~4))

merged_main <- main_data %>%
  left_join(record_data[, c("Year", "Sex", "CR", "Race_code")],
            by = c("Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D." = "Race_code",
                   "Year" = "Year",
                   "Sex..0.F..1.M." = "Sex")) %>%
  dplyr::rename(Race_code = Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D.,
         Sex = Sex..0.F..1.M.,
         Age = Age..yr.,
         SR = SR.W.m2) %>%
  mutate(Gender = factor(ifelse(Sex == 0, "Female", "Male"))) %>%
  mutate(Marathon = case_when(Race_code == 0 ~ "Boston",
                          Race_code == 1~ "NYC",
                          Race_code == 2 ~ "Chicago",
```

```
                          Race_code == 3 ~ "Twin Cities",
                          Race_code == 4 ~ "Grandmas")) %>%
  mutate(Flag = factor(Flag,
                       levels = c("White", "Green", "Yellow", "Red", "Black")))  %>%
  mutate(NetRaceTime = as.numeric(as.difftime(CR, units = "mins") * (1+X.CR/100))) %>%
  mutate(Age_Group = cut(Age, breaks = c(0, 14, 19, 29, 39, 49, 59, 69, 79, 92),
                         labels = c("<= 14", "15-19", "20-29", "30-39",
                                    "40-49", "50-59", "60-69", "70-79", ">= 80"),
                         right = TRUE))


#Clean up AQI by CBSA
aqi_data = aqi_data %>%
  distinct()

AP_mean = aqi_data %>%
  group_by(marathon, date_local, parameter, sample_duration) %>%
  summarise(daily_mean = mean(arithmetic_mean, na.rm = TRUE)) %>%
  mutate(parameter_duration = paste0(parameter, "-", sample_duration))


AP_pivot = AP_mean[,c("marathon", "date_local", "parameter_duration", "daily_mean")] %>%
  pivot_wider(names_from = parameter_duration, values_from = daily_mean) %>%
  mutate(Year = year(date_local)) %>%
  mutate(PM2.5 = coalesce(`PM2.5 - Local Conditions-24 HOUR`,
                      `PM2.5 - Local Conditions-24-HR BLK AVG`))%>%
  dplyr::select(
    "marathon",
    "date_local",
    "Year",
    "Sulfur dioxide-1 HOUR",
    "Ozone-1 HOUR",
    "Nitrogen dioxide (NO2)-1 HOUR",
    "PM2.5"
  ) %>%
  rename("SO2" = "Sulfur dioxide-1 HOUR",
         "NO2" = "Nitrogen dioxide (NO2)-1 HOUR",
         "Ozone" = "Ozone-1 HOUR")

merged_main = merged_main %>%
  left_join(AP_pivot,
            by = c("Marathon" = "marathon",
                   "Year" = "Year")) %>%
  mutate(Wind_s = scale(Wind),
         WBGT_s = scale(WBGT),
         SR_s = scale(SR),
         X.rh_s = scale(X.rh),
         Ozone_s = scale(Ozone),
         PM2.5_s = scale(PM2.5),
         SO2_s = scale(SO2),
         NO2_s = scale(NO2)
         ) %>%
  mutate(log_X.CR = log(X.CR),
```

```r
         log_NetRaceTime = log(NetRaceTime))

# For course records and environmental parameters only
CR_merged = merged_main  %>%
  dplyr::select(Marathon, CR, Gender, WBGT, Flag, Wind,
    X.rh, SR, NO2, SO2, Ozone, PM2.5, WBGT_s, Wind_s,
    X.rh_s, SR_s, NO2_s, SO2_s, Ozone_s, PM2.5_s) %>%
  distinct() %>%
  mutate(ChipTime = as.numeric(as.difftime(CR, units = "mins")))
# Baseline summary
merged_main %>%
  select(
    Marathon,
    Gender,
    Age
  ) %>%
  tbl_summary(
    statistic = list(
      all_continuous() ~ "{mean} ({sd})",
      all_categorical() ~ "{n} ({p}%)"
    ),
    by = Marathon,
    digits = all_continuous() ~ 2,
    missing = "no",
    type = list(
      Gender ~ "categorical"
    )
  ) %>%
  add_p() %>%
  modify_caption(caption = "Participant Characteristics by Marathon Race, N = {N}") %>%
  as_kable_extra(
    booktabs = TRUE,
    longtable = TRUE,
    linesep = ""
  ) %>%
  kableExtra::kable_styling(
    position = "center",
    latex_options = c("scale_down", "striped", "repeat_header"),
    stripe_color = "gray!15",
    font_size = 10
  )

dist_plot <- ggplot(merged_main %>% filter(Age >= 15), aes(x = X.CR, fill = Gender)) +
  geom_histogram(
    position = "identity",
    binwidth = 10,
    alpha = 0.6,
    color = NA
  ) +
  scale_fill_manual(values = c("Female" = "lightcoral", "Male" = "lightblue")) +
  facet_wrap( ~ Age_Group, scales = "free_y", nrow = 2) +
  theme_minimal() +
  labs(title = "Figure 1: Marathon Performance Distribution by Gender and Age Group",
```

```r
      x = "Percent off current course record", y = "Count") +
  theme(
    strip.text = element_text(face = "bold", size = 14),
    axis.text.x = element_text(size = 12),
    axis.text.y = element_text(size = 12),
    plot.title = element_text(hjust = 0.5, size = 16),
    axis.title.x = element_text(size = 14),
    axis.title.y = element_text(size = 14),
    legend.title = element_blank(),
    legend.position = "bottom"
  )
dist_plot
merged_main  %>%
  dplyr::select(
    Marathon,
    WBGT,
    Flag,
    Td..C,
    Tw..C,
    Tg..C,
    DP,
    Wind,
    X.rh,
    SR,
    NO2,
    SO2,
    Ozone,
    PM2.5
  ) %>%
  distinct() %>%
  tbl_summary(
    statistic = list(all_continuous() ~ c("{mean} ({sd})", "{min}, {max}"),
                     all_categorical() ~ "{n} ({p}%)"),
    by = Marathon,
    digits = all_continuous() ~ 1,
    missing = "no",
    type = list(
      WBGT ~ "continuous2",
      Flag ~ "categorical",
      Td..C ~ "continuous2",
      Tw..C ~ "continuous2",
      Tg..C ~ "continuous2",
      DP ~ "continuous2",
      X.rh ~ "continuous2",
      SR ~ "continuous2",
      Wind ~"continuous2",
      SO2  ~"continuous2",
      NO2 ~"continuous2",
      Ozone ~"continuous2",
      PM2.5 ~"continuous2"
    ),
    label = list(
      WBGT = "WBGT (°C)",
```

```
      Flag = "WBGT flag",
      Td..C = "Td (°C)",
      Tw..C = "Tw (°C)",
      Tg..C = "Tg (°C)",
      DP = "DP (°C)",
      X.rh = "% relative humidity",
      Wind = "Wind speed (Km/hr)",
      SR = "SR (W/m^2)",
      NO2 = "NO2 (ppb)",
      SO2 = "SO2 (ppb)",
      Ozone = "O3 (ppm)",
      PM2.5 = "PM2.5 (microgram/m^3)"
    )
  ) %>% add_p() %>%
   modify_caption(caption = "Summary of Environmental Factors by Marathon, N = {N}") %>%
  as_kable_extra(
    booktabs = TRUE,
    longtable = TRUE,
    linesep = "",
    format = "latex"
  ) %>%
  kableExtra::kable_styling(
    position = "center",
    latex_options = c("striped", "repeat_header"),
    stripe_color = "gray!15",
    font_size = 8
  )
#, out.width= "65%", out.extra='style="float:right; padding:10px"'
all_environ_factors = c("SO2", "NO2", "Ozone", "PM2.5", "WBGT", "Wind", "DP",
                        "Td..C", "Tw..C", "X.rh", "Tg..C", "SR")
environ_data <- merged_main[, all_environ_factors]

custom_labels <- c( "SO2 (ppb)", "NO2 (ppb)",
                    "Ozone (ppm)", "PM2.5 (microgram/m^3)",
                    "WBGT (°C)", "Wind Speed (km/hr)", "DP (°C)",
                   "Td (°C)", "Tw (°C)",
                   "%rh", "Tg (°C)", "SR (W/m^2)")


cor_matrix <- cor(environ_data, use = "complete.obs", method = "spearman")
colnames(cor_matrix) <- custom_labels
rownames(cor_matrix) <- custom_labels

corrplot::corrplot(cor_matrix,
                   type = "lower",
                   method = "color",
                   tl.cex = 0.6,
                   tl.labels = custom_labels,
                   tl.col = "black",
                   col = colorRampPalette(c("darkblue", "beige", "brown"))(200),
                   addCoef.col = "black",
                   number.cex = 0.5,
                   number.font = 1.3
```

```r
)
title("Figure 2: Correlation between Environmental Parameters", cex.main = 0.8)
distinct_environ = CR_merged[, c("WBGT", "X.rh", "Wind","SR",
                                 "NO2", "SO2", "Ozone", "PM2.5")] %>%
  distinct()

custom_labels <- list(
  "X.rh" = "% relative humidity",
  "Wind" = "Wind Speed",
  "WBGT" = "WBGT",
  "SR" = "SR",
  "NO2" = "NO2",
  "SO2" = "SO2",
  "Ozone" = "Ozone",
  "PM2.5" = "PM2.5"
)

# Define a wrapper function to create a histogram for each environmental factor
create_histogram <- function(df, var_name) {
  missing_count <- sum(is.na(df[[var_name]]))
  total_count <- nrow(df)
  missing_percent <- (missing_count / total_count) * 100

  label <- custom_labels[[var_name]] %||% var_name

  ggplot(df, aes(x = .data[[var_name]])) +
    geom_histogram(bins = 10, fill = "lightblue", alpha = 0.7) +
    labs(title = label, x = "",
         y = ifelse(var_name %in% c("WBGT", "NO2"), "Count", ""))+
      annotate(
        "text",
        x = Inf,
        y = Inf,
        label = paste0(
          "Missing: ", missing_count, "\n(",
          round(missing_percent, 2), "%)"
        ),
        hjust = 1.1,
        vjust = 1.6,
        size = 5,
        color = "#CC3333",
        fontface = "italic"
      ) +
    theme_minimal() +
    theme(
      plot.title = element_text(hjust = 0.5, size = 14),
      axis.title.x = element_text(size = 12),
      axis.title.y = element_text(size = ifelse(var_name %in% c("WBGT", "NO2"), 12, 0)),
      axis.text.y = element_text(size = ifelse(var_name %in% c("WBGT", "NO2"), 12, 0))
    )
}

environ_factors <- c("WBGT", "X.rh", "Wind", "SR", "NO2", "SO2", "Ozone", "PM2.5")
```

```r
histograms1 <- lapply(environ_factors[1:4],
                      function(var) create_histogram(distinct_environ, var))

combined_plot1 <- wrap_plots(histograms1, ncol = 4,  scales = "free_x")+
  plot_annotation(
    title = "Figure 3: Distribution of Environmental Factors with Missingness",
    theme = theme(plot.title = element_text(hjust = 0.5, size = 20))
  )

print(combined_plot1)

histograms2 <- lapply(environ_factors[5:8],
                      function(var) create_histogram(distinct_environ, var))

combined_plot2 <- wrap_plots(histograms2, ncol = 4,  scales = "free_x")

print(combined_plot2)

off_record_gender_age = ggplot(merged_main,
                               aes(x = Age, y = X.CR, color = Gender)) +
  # Loess smoothing with 95% CI
   geom_smooth(method = "loess", se = TRUE, size = 0.5) +
  labs(
    title = "Percent off Current Course Record vs. Age",
    x = "Age (yrs)",
    y = "Percent off current course record "
  ) +
  scale_y_continuous(limits = c(0, 300)) +
  scale_x_continuous(breaks = seq(10, 100, by = 10)) +
   scale_color_manual(values = c("Male" = "steelblue", "Female" = "darkred")) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 8),
    axis.title.x = element_text(size = 8),
    axis.title.y = element_text(size = 8),
    legend.title = element_blank(),
    legend.position = "none"
  )

race_time_gender_age = ggplot(merged_main, aes(x = Age,
                                               y = NetRaceTime,
                                               color = Gender)) +
   geom_smooth(method = "loess", se = TRUE, size = 0.5) +
  labs(
    title = "Net Race Time vs. Age",
    x = "Age (yrs)",
    y = "Net Race time (minutes)",
    legend = "Gender"
  ) +
  scale_x_continuous(breaks = seq(10, 100, by = 10)) +
   scale_color_manual(values = c("Male" = "steelblue", "Female" = "darkred")) +
  theme_minimal() +
  theme(
```

```
    plot.title = element_text(hjust = 0.5, face = "bold", size = 8),
    axis.title.x = element_text(size = 8),
    axis.title.y = element_text(size = 8),
    legend.title = element_text(size = 7),
    legend.text = element_text(size = 6),
    legend.position = "right"
  )

combined_plot <- wrap_plots(off_record_gender_age, race_time_gender_age, ncol = 2) +
  plot_annotation(
    title = "Figure 4: Impact of Age on Marathon Performance by Gender and WBGT Flag",
    theme = theme(plot.title = element_text(hjust = 0.5, size = 12))
  ) +
  plot_layout(ncol = 2, widths = c(1, 1.15)) &
  theme(plot.margin = unit(c(0.1, 0.1, 0.1, 0.1), "cm"))

combined_plot
off_record_WBGT_age = ggplot(merged_main %>% filter(!is.na(Flag)),
                              aes(x = Age, y = X.CR, color = Flag)) +
  # Loess smoothing with 95% CI
   geom_smooth(method = "loess", se = TRUE, size = 0.5) +
  labs(
    x = "Age (yrs)",
    y = "Percent off current course record "
  ) +
  scale_y_continuous(limits = c(0, 300)) +
  scale_x_continuous(breaks = seq(10, 100, by = 10)) +
  scale_color_manual(values = c("White" = "lightblue", "Green" = "yellowgreen",
                                "Yellow" = "orange", "Red" = "lightcoral")) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 8),
    axis.title.x = element_text(size = 8),
    axis.title.y = element_text(size = 8),
    legend.title = element_blank(),
    legend.position = "none"
  )

race_time_WBGT_age = ggplot(merged_main%>% filter(!is.na(Flag)),
                           aes(x = Age, y = NetRaceTime, color = Flag)) +
   geom_smooth(method = "loess", se = TRUE, size = 0.5) +
  labs(
    x = "Age (yrs)",
    y = "Net Race time (minutes)",
    legend = "WBGT Flag"
  ) +
  scale_x_continuous(breaks = seq(10, 100, by = 10)) +
  scale_color_manual(values = c("White" = "lightblue", "Green" = "yellowgreen",
                                "Yellow" = "orange", "Red" = "lightcoral")) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", size = 8),
    axis.title.x = element_text(size = 8),
```

```r
    axis.title.y = element_text(size = 8),
    legend.title = element_text(size = 7),
    legend.text = element_text(size = 6),
    legend.position = "right"
  )

combined_plot <- wrap_plots(off_record_WBGT_age, race_time_WBGT_age, ncol = 2) +
  plot_layout(ncol = 2, widths = c(1, 1.15)) &
  theme(plot.margin = unit(c(0.1, 0.1, 0.1, 0.1), "cm"))

combined_plot
cor_table <- data.frame(
  Environmental_Factor = environ_factors,
  NetRaceTime_Corr = NA,
  ChipTime_Corr = NA
)


for (i in seq_along(environ_factors)) {
  factor <- environ_factors[i]

  # Correlation for NetRaceTime for all runners
  cor_table$NetRaceTime_Corr[i] <- round(cor(merged_main[[factor]],
                                       merged_main$NetRaceTime,
                                       use = "complete.obs",
                                       method = "spearman"), 4)

  # Correlation for ChipTime in records
  cor_table$ChipTime_Corr[i] <- round(cor(CR_merged[[factor]],
                                    CR_merged$ChipTime,
                                    use = "complete.obs",
                                    method = "spearman"), 4)
}

colnames(cor_table)[2:3] <- c("Net race time in all runners", "Course record")
rownames(cor_table) <- c("WBGT", "%rh", "Wind Speed", "Solar Radiation",
                         "NO2", "SO2", "Ozone", "PM2.5")
cor_table[, -1] %>%
  kable(caption = "Spearman Correlation between Environmental Factors and Net Race Time
        for General Runners and Record Setters") %>%
    kableExtra::kable_styling(
    position = "center",
    latex_options = c("scale_down", "striped", "repeat_header"),
    stripe_color = "gray!15",
    font_size = 11
  )

# List of pollutant variables and their labels
weather <- c("Wind", "WBGT", "SR", "X.rh")
weather_titles <- c("Wind speed", "Wet Bulb Globe Temperature",
            "Solar radiation ", "Percent relative humidity")
units <- c(" (Km/hr)", " (°C)", " (W/m^2)", "")
```

```r
# Create a list of ggplot objects
plots <- lapply(seq_along(weather), function(i) {
  ggplot(merged_main, aes_string(x = weather[i], y = "X.CR" , color = "Gender")) +
    geom_smooth(method = "loess", se = TRUE, size = 1) +
    labs(
      title = weather_titles[i],
      x = paste0(weather[i], " ", units[i]),
      # Only show y-axis label on the first plot
      y = if (i == 1) "Best Time (%CR)" else NULL
    ) +
    scale_color_manual(values = c("Male" = "steelblue", "Female" = "darkred")) +
    theme_minimal() +
    theme(
      plot.title = element_text(hjust = 0.5, size = 14),
      axis.title.x = element_text(size = 14),
      # Hide y-axis title and text for non-row-leading plots
      axis.title.y = element_text(size = if (i == 1) 14 else 0),
      axis.text.y = element_text(size = if (i == 1) 12 else 0),
      legend.position = if (i == 4) "right" else "none"
    )
})

# Combine the four plots into a single row with shared y-axis
combined_plot <- wrap_plots(plots, ncol = 4) &
  theme(plot.margin = unit(c(0.1, 0.1, 0.1, 0.1), "cm"))

combined_plot <- combined_plot +
  plot_annotation(
    title = "Figure 5: Impact of Environmental Factors on Marathon Performance by Gender using LOESS",
    theme = theme(plot.title = element_text(hjust = 0.5, size = 18))
  )

combined_plot
# List of pollutant variables and their labels
pollutants <- c("Ozone", "NO2", "SO2", "PM2.5")
titles <- c("Ozone", "NO2", "SO2", "PM2.5")
units <- c(" (ppm)", " (ppb)", " (ppb)", "(microgram/m^3)")

# Create a list of ggplot objects
plots <- lapply(seq_along(pollutants), function(i) {
  ggplot(merged_main, aes_string(x = pollutants[i], y = "X.CR" , color = "Gender")) +
    geom_smooth(method = "loess", se = TRUE, size = 1) +
    labs(
      title = titles[i],
      x = paste0(pollutants[i], " ", units[i]),
      # Only show y-axis label on the first plot
      y = if (i == 1) "Best Time (%CR)" else NULL
    ) +
    scale_color_manual(values = c("Male" = "steelblue", "Female" = "darkred")) +
    theme_minimal() +
    theme(
      plot.title = element_text(hjust = 0.5, size = 14),
      axis.title.x = element_text(size = 14),
```

```r
      # Hide y-axis title and text for non-row-leading plots
      axis.title.y = element_text(size = if (i == 1) 14 else 0),
      axis.text.y = element_text(size = if (i == 1) 12 else 0),
      legend.position = if (i == 4) "right" else "none"
    )
})

# Combine the four plots into a single row with shared y-axis
combined_plot <- wrap_plots(plots, ncol = 4) &
  theme(plot.margin = unit(c(0.1, 0.1, 0.1, 0.1), "cm"))
combined_plot
ggplot(merged_main %>% filter(Age >= 15), aes(x = WBGT, y = X.CR, color = Gender)) +
  geom_point(size = 0.3, alpha = 0.3) +
  geom_smooth(method = "loess", se = TRUE, size = 0.75, color = "black") +
  facet_grid(Gender~Age_Group) +
  scale_color_manual(values = c("Male" = "steelblue", "Female" = "darkred")) +
  scale_y_continuous(limits = c(0, 300)) +
  labs(title = "Figure 6: the Impact of WBGT and NO2 on Marathon Performance for General Runners",
       x = "WBGT (°C)", y = "Percent off current course record") +
  theme_minimal() +
  theme(legend.position = "none",
        plot.title = element_text(size = 14, hjust = 0.5))
ggplot(merged_main %>% filter(Age >= 15), aes(x = NO2, y = X.CR, color = Gender)) +
  geom_point(size = 0.3, alpha = 0.3) +
  geom_smooth(method = "loess", se = TRUE, size = 0.75, color = "black") +
  facet_grid(Gender~Age_Group) +
  scale_color_manual(values = c("Male" = "steelblue", "Female" = "darkred")) +
  scale_y_continuous(limits = c(0, 300)) +
  labs(x = "NO2 (ppm)", y = "Percent off current course record") +
  theme_minimal() +
  theme(legend.position = "none")

std_environ_factors = c("Wind_s", "WBGT_s", "SR_s", "X.rh_s",
                        "Ozone_s", "SO2_s", "NO2_s", "PM2.5_s")

std_x_labels <- list(
  "X.rh_s" = "% relative humidity",
  "Wind_s" = "Wind Speed",
  "WBGT_s" = "WBGT",
  "SR_s" = "SR",
  "NO2_s" = "NO2",
  "SO2_s" = "SO2",
  "Ozone_s" = "Ozone",
  "PM2.5_s" = "PM2.5"
)

# Wrapper funct to fit GAM, create plot, and extract adjusted R-squared
create_gam_plot <- function(df, show_y_axis = TRUE, x_var,
                            y_var = "X.CR", ylab_text = "% off course record") {

  # Filter out missing and non-finite values
  model_data <- df %>%
    filter(.data[[y_var]] > 0) %>%
```

```r
    filter(!is.na(.data[[x_var]]) & !is.na(.data[[y_var]]) &
             is.finite(.data[[x_var]]) & is.finite(.data[[y_var]]))

  gam_fit <- gam(as.formula(paste(y_var, "~ s(", x_var, ")")),
                 data = model_data,
                 family = Gamma(link = "log"))

  adj_r2 <- summary(gam_fit)$r.sq

  x_label <- paste0("Standardized ", std_x_labels[[x_var]] %||% x_var)
  ylab_text <- if (show_y_axis) ylab_text else ""

  mar_setting <- if (show_y_axis) c(5, 4, 4, 2) else c(5, 1.5, 4, 2)
  par(mar = mar_setting)

  plot(gam_fit, se = TRUE, shade = TRUE, rug = FALSE,
       xlab = " ", ylab = ylab_text, main = x_label,
       cex.lab = if (show_y_axis) 1.5 else 0.1, cex.main = 1.5)

  # Add annotation for adjusted R²
  legend(
    "topright",
    legend = paste0("Adj R² = ", 100*round(adj_r2, 3), "%"),
    bty = "n",
    text.col = "#CC3333",
    cex = 1.75
  )
}

# Create all plots
# Set up a 2x4 plot layout with outer margins for title
par(mfrow = c(1, 4), oma = c(0, 0, 3, 0))
for (i in 1:4) {
  show_y_axis <- (i %% 4 == 1)  # Show y-axis only for the first plot in each row
  create_gam_plot(merged_main, x_var = std_environ_factors[i],
                  y_var = "X.CR", show_y_axis = show_y_axis,
                  ylab_text = "% off course record")
}
mtext("Figure 7: GAM Fits of Environmental Factors on % off Course Record for General Runners",
      outer = TRUE, cex = 1.5, font = 2)
par(mfrow = c(1, 4))
for (i in 5:8) {
  show_y_axis <- (i %% 4 == 1)  # Show y-axis only for the first plot in each row
  create_gam_plot(merged_main, x_var = std_environ_factors[i],
                  y_var = "X.CR", show_y_axis = show_y_axis,
                  ylab_text = "% off course record")
}
# Data frame to store results
gam_results <- data.frame(
  Variable = character(),
  n = numeric(),
  F_value = numeric(),
  P_value = numeric(),
```

```r
  Adj_R2 = numeric(),
  Deviance_Explained = numeric(),
  stringsAsFactors = FALSE
)


get_gam_info <- function(df, x_var, y_var = "X.CR") {

  # Filter data for valid values
  model_data <- df %>%
    filter(.data[[y_var]] > 0) %>%
    filter(!is.na(.data[[x_var]]) & !is.na(.data[[y_var]]) &
             is.finite(.data[[x_var]]) & is.finite(.data[[y_var]]))

  gam_fit <- gam(as.formula(paste0(y_var, " ~ s(", x_var, ")")),
                 data = model_data, family = Gamma(link = "log"))

  gam_summary <- summary(gam_fit)
  n <- gam_summary$n

  # Extracting approximate F-value and p-value for the smooth term
  smooth_terms <- gam_summary$s.table

  F_value <- smooth_terms[1, "F"]
  p_value <- smooth_terms[1, "p-value"]

  # Adjusted R-squared and deviance explained
  adj_r2 <- paste0(round(gam_summary$r.sq, 4) * 100, "%")
  deviance_explained <- paste0(round(gam_summary$dev.expl, 4)*100, "%")

  return(
    list(
      n = n,
      F_value = round(F_value, 4),
      P_value = round(p_value, 4),
      Adj_R2 = adj_r2,
      Deviance_Explained = deviance_explained
    )
  )
}

for (x_var in std_environ_factors) {
  gam_info <- get_gam_info(merged_main, x_var)

  gam_results <- rbind(gam_results, data.frame(
    Variable = std_x_labels[[x_var]] %||% x_var,
    n = gam_info$n,
    F_value = gam_info$F_value,
    Approx_P_value = gam_info$P_value,
    Adj_R2 = gam_info$Adj_R2,
    Deviance_Explained = gam_info$Deviance_Explained
  ))
}
```

```r
colnames(gam_results) <- c("Standardized variable", "n","Approxi. F-value",
                           "Approxi. p-value", "Adj. R2", "% Dev. explained")

gam_results %>%
  kable(caption = "GAM Results- Percent off Course Record for General Runners vs. Standardized Environme
        align = "c") %>%
  kable_styling(full_width = FALSE,
                latex_options = c("striped", "repeat_header"),
                stripe_color = "gray!15",
                font_size = 11,
                position = "center")

# List of pollutant variables and their labels
weather <- c("Wind", "WBGT", "SR", "X.rh")
weather_titles <- c("Wind speed", "Wet Bulb Globe Temperature",
            "Solar radiation ", "Percent relative humidity")
units <- c(" (Km/hr)", " (°C)", " (W/m^2)", "")

# Create a list of ggplot objects
plots <- lapply(seq_along(weather), function(i) {
  ggplot(CR_merged, aes_string(x = weather[i], y = "ChipTime" , color = "Gender")) +
    geom_smooth(method = "loess", se = TRUE, size = 1) +
    labs(
      title = weather_titles[i],
      x = paste0(weather[i], " ", units[i]),
      # Only show y-axis label on the first plot
      y = if (i == 1) "Course record (minute)" else NULL
    ) +
    scale_color_manual(values = c("Male" = "steelblue", "Female" = "darkred")) +
    theme_minimal() +
    theme(
      plot.title = element_text(hjust = 0.5, size = 14),
      axis.title.x = element_text(size = 14),
      # Hide y-axis title and text for non-row-leading plots
      axis.title.y = element_text(size = if (i == 1) 14 else 0),
      axis.text.y = element_text(size = if (i == 1) 12 else 0),
      legend.position = if (i == 4) "right" else "none"
    )
})

# Combine the four plots into a single row with shared y-axis
combined_plot <- wrap_plots(plots, ncol = 4) &
  theme(plot.margin = unit(c(0.1, 0.1, 0.1, 0.1), "cm"))

combined_plot <- combined_plot +
  plot_annotation(
    title = "Figure 8: Impact of Environmental Factors on Course Records by Gender using LOESS",
    theme = theme(plot.title = element_text(hjust = 0.5, size = 18))
  )

combined_plot
# List of pollutant variables and their labels
pollutants <- c("Ozone", "NO2", "SO2", "PM2.5")
```

```r
titles <- c("Ozone", "NO2", "SO2", "PM2.5")
units <- c(" (ppm)", " (ppb)", " (ppb)", "(microgram/m^3)")

# Create a list of ggplot objects
plots <- lapply(seq_along(pollutants), function(i) {
  ggplot(CR_merged, aes_string(x = pollutants[i], y = "ChipTime" , color = "Gender")) +
    geom_smooth(method = "loess", se = TRUE, size = 1) +
    labs(
      title = titles[i],
      x = paste0(pollutants[i], " ", units[i]),
      # Only show y-axis label on the first plot
      y = if (i == 1) "Course record (minute)" else NULL
    ) +
    scale_color_manual(values = c("Male" = "steelblue", "Female" = "darkred")) +
    theme_minimal() +
    theme(
      plot.title = element_text(hjust = 0.5, size = 14),
      axis.title.x = element_text(size = 14),
      # Hide y-axis title and text for non-row-leading plots
      axis.title.y = element_text(size = if (i == 1) 14 else 0),
      axis.text.y = element_text(size = if (i == 1) 12 else 0),
      legend.position = if (i == 4) "right" else "none"
    )
})

# Combine the four plots into a single row with shared y-axis
combined_plot <- wrap_plots(plots, ncol = 4) &
  theme(plot.margin = unit(c(0.1, 0.1, 0.1, 0.1), "cm"))
combined_plot
# Wrapper func to fit lm and extract summary statistics for the environmental variable
fit_lm <- function(data, var) {
  model <- lm(as.formula(paste("log(ChipTime) ~", var, "+ Gender")), data = data)
  model_summary <- summary(model)

  # Extract beta, CI, p-value, and adjusted R-squared for the environmental variable
  beta <- coef(model)[var]
  ci <- confint(model)[var, ]
  p_value <- coef(summary(model))[var, "Pr(>|t|)"]
  adj_r2 <- model_summary$adj.r.squared

  return(data.frame(Variable = var,
                    Beta = beta,
                    CI_lower = ci[1],
                    CI_upper = ci[2],
                    P_value = p_value,
                    Adj_R2 = adj_r2))
}


results <- list()

# generate gender only R2
gender_model <- lm(log(ChipTime) ~ Gender, data = CR_merged)
```

```r
gender_summary <- summary(gender_model)
gender_adj_r2 <- gender_summary$adj.r.squared


# Loop through the environmental variables
for (var in std_environ_factors) {
  results[[var]] <- fit_lm(CR_merged, var)
}

results_df <- bind_rows(results)
rownames(results_df) <- c("WBGT", "%rh", "Wind Speed", "Solar Radiation",
                          "NO2", "SO2", "Ozone", "PM2.5")

# Format the table
results_df[,-1] %>%
  mutate(P_value = round(P_value, 4),
         Adj_R2 = paste0(round(Adj_R2- gender_adj_r2, 4)*100, "%"),
         Beta = round(Beta, 4),
         CI_lower = round(CI_lower, 4),
         CI_upper = round(CI_upper, 4)) %>%
  kable(caption = "Linear Model Results- log(Course Records) vs. Environmental Factors (Controlled for G
        col.names = c("Standardized Var.", "Beta", "CI Lower", "CI Upper", "p-value", "Approxi. Isolated
        linesep = c("","\\addlinespace"),
        align = "c") %>%
  kable_styling(full_width = FALSE,
                latex_options = c("striped", "repeat_header"),
                stripe_color = "gray!15",
                font_size = 11,
                position = "center")

# Impact of environmental factors on performance by age and gender

ggplot(merged_main %>% filter(Age >= 15), aes(x = SR, y = X.CR, color = Gender)) +
  geom_point(size = 0.3, alpha = 0.3) +
  geom_smooth(method = "loess", se = TRUE, size = 0.75, color = "black") +
  facet_grid(Gender~Age_Group) +
  scale_color_manual(values = c("Male" = "steelblue", "Female" = "darkred")) +
  scale_y_continuous(limits = c(0, 300)) +
  labs(title = "Figure S1: The Impact of Solar Radiation on Marathon Performance by Age and Gender",
    x = "SR (W/m^2)", y = "Percent off current course record") +
  theme_minimal() +
  theme(legend.position = "none")
ggplot(merged_main %>% filter(Age >= 15), aes(x = X.rh, y = X.CR, color = Gender)) +
  geom_point(size = 0.3, alpha = 0.3) +
  geom_smooth(method = "loess", se = TRUE, size = 0.75, color = "black") +
  facet_grid(Gender~Age_Group) +
  scale_color_manual(values = c("Male" = "steelblue", "Female" = "darkred")) +
  scale_y_continuous(limits = c(0, 300)) +
  labs(title = "Figure S2: The Impact of Percent Relative Humidity on Marathon Performance by Age and Ge
    x = "%rh", y = "Percent off current course record") +
  theme_minimal() +
  theme(legend.position = "none")
ggplot(merged_main %>% filter(Age >= 15), aes(x = Wind, y = X.CR, color = Gender)) +
```

```r
  geom_point(size = 0.3, alpha = 0.3) +
  geom_smooth(method = "loess", se = TRUE, size = 0.75, color = "black") +
  facet_grid(Gender~Age_Group) +
  scale_color_manual(values = c("Male" = "steelblue", "Female" = "darkred")) +
  scale_y_continuous(limits = c(0, 300)) +
  labs(title = "Figure S3: The Impact of Wind Speed on Marathon Performance by Age and Gender",
    x = "Wind speed (km/hr)", y = "Percent off current course record") +
  theme_minimal() +
  theme(legend.position = "none")
ggplot(merged_main %>% filter(Age >= 15), aes(x = SO2, y = X.CR, color = Gender)) +
  geom_point(size = 0.3, alpha = 0.3) +
  geom_smooth(method = "loess", se = TRUE, size = 0.75, color = "black") +
  facet_grid(Gender~Age_Group) +
  scale_color_manual(values = c("Male" = "steelblue", "Female" = "darkred")) +
  scale_y_continuous(limits = c(0, 300)) +
  labs(title = "Figure S4: The Impact of SO2 on Marathon Performance by Age and Gender",
    x = "SO2 (ppm)", y = "Percent off current course record") +
  theme_minimal() +
  theme(legend.position = "none")
ggplot(merged_main %>% filter(Age >= 15), aes(x = Ozone_s, y = X.CR, color = Gender)) +
  geom_point(size = 0.3, alpha = 0.3) +
  geom_smooth(method = "loess", se = TRUE, size = 0.75, color = "black") +
  facet_grid(Gender~Age_Group) +
  scale_color_manual(values = c("Male" = "steelblue", "Female" = "darkred")) +
  scale_y_continuous(limits = c(0, 300)) +
  labs(title = "Figure S5: The Impact of Ozone on Marathon Performance by Age and Gender",
    x = "Ozone (ppb)", y = "Percent off current course record") +
  theme_minimal() +
  theme(legend.position = "none")
ggplot(merged_main %>% filter(Age >= 15), aes(x = PM2.5, y = X.CR, color = Gender)) +
  geom_point(size = 0.3, alpha = 0.3) +
  geom_smooth(method = "loess", se = TRUE, size = 0.75, color = "black") +
  facet_grid(Gender~Age_Group) +
  scale_color_manual(values = c("Male" = "steelblue", "Female" = "darkred")) +
  scale_y_continuous(limits = c(0, 300)) +
  labs(title = "Figure S6: The Impact of PM2.5 on Marathon Performance by Age and Gender",
    x = "PM2.5 (microgram/m^3)", y = "Percent off current course record") +
  theme_minimal() +
  theme(legend.position = "none")
```