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, which represent the Boston marathon, Chicago marathon, New York City marathon, Twin Cities marathon and Grandma's marathon respectively. The Year variable was also transformed into a factor with levels from 1993 to 2016 for more precise control of the later plots. 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.

In order to include more accurate weather information, the Air Quality Index calculated using ozone was also added to the data for analysis. Specifically, we calculated the average AQI values of multiple stations in each city during each game as the final AQI for analysis.

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. In addition, Figure 1a shows that there was scale error in the relative humanity variable, with some records showing that the relative humanity is less than 1. We multiplied the data for this portion of relative humanity by 100 to maintain data consistency (Figure 1b). The summary of environmental variables is shown in Table 1, and the summary of personal information is shown in Table 2.

According to Table 1, it can be observed that there are there is no race in NYC for WBGT $>$ 28. And observing Table 2, we found that in all cases, the gender ratio of participants in the competition is the same. And it can be preliminarily assumed that as the temperature rises, the performance of runners decreases.

Figure 1a: Histogram of relative humidity

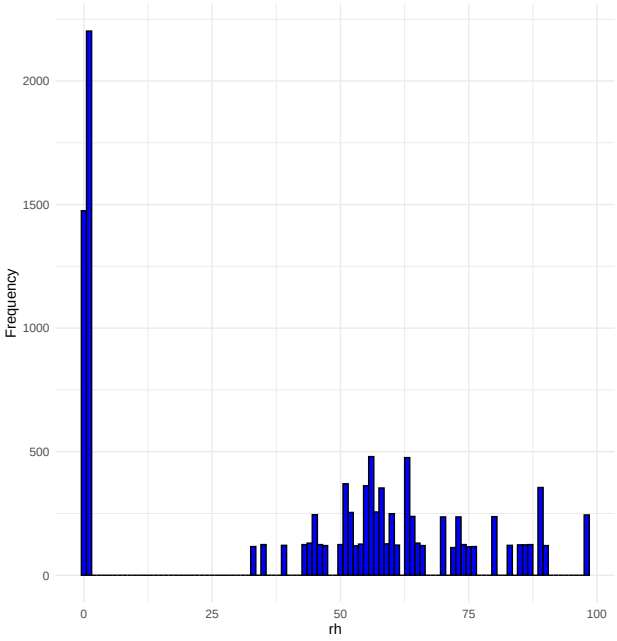


Figure 1b: Histogram of corrected relative humidity

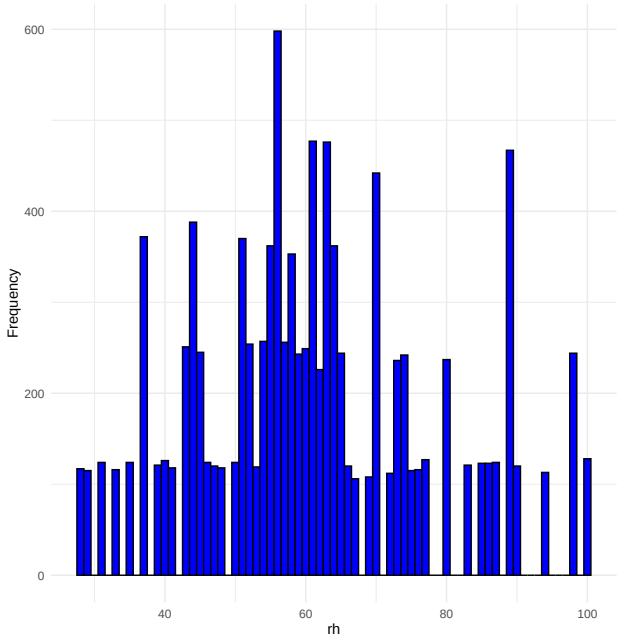


Table 1: Summary Statistics of Environmental Information by Flag

Characteristic	**White** N = 3,753	**Green** N = 4,706	**Yellow** N = 2,022	**Red** N = 592
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	56.47	62.34	64.73	66.63
Min, Max	30.72, 98.25	27.81, 98.33	29.27, 100.00	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

Table 2: Summary Statistics of Personal Information 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%)

Table 2: Summary Statistics of Personal Information by Flag *(continued)*

Characteristic	**White** N = 3,753	**Green** N = 4,706	**Yellow** N = 2,022	**Red** N = 592
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%)
Sex				
Female	1,769 (47%)	2,222 (47%)	948 (47%)	279 (47%)
Male	1,984 (53%)	2,484 (53%)	1,074 (53%)	313 (53%)
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

[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

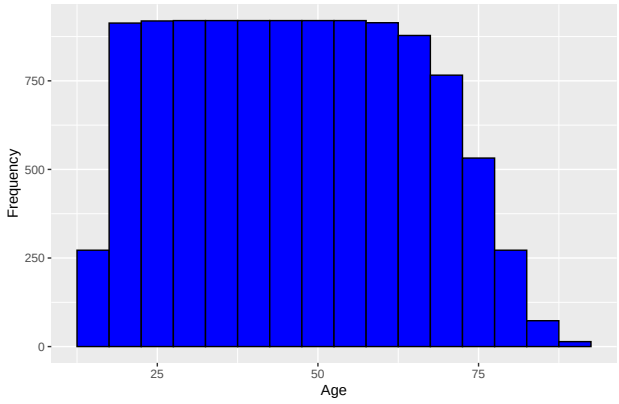


Figure 2b. Distribution of Age

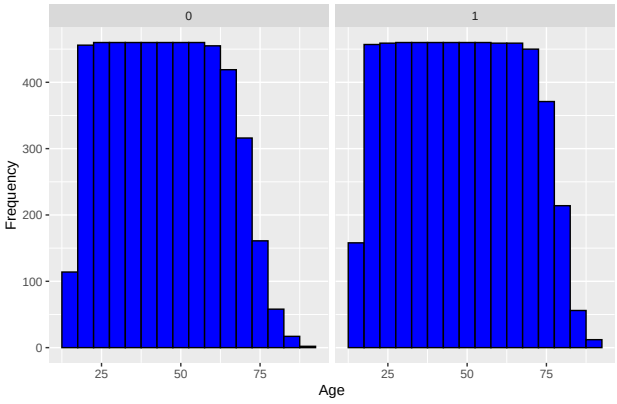


Figure 2c. Distribution of Percent off

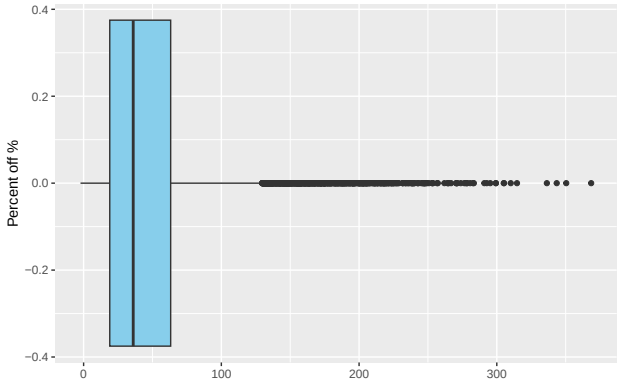
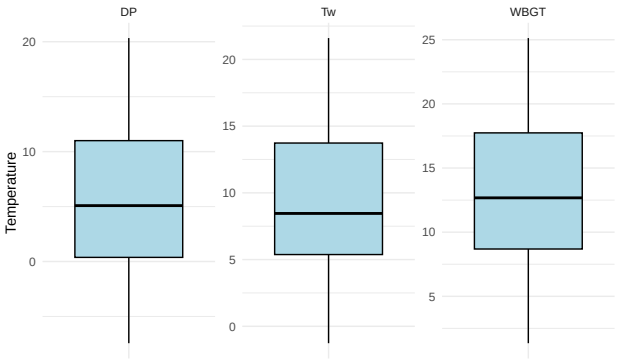


Figure 2d. Distribution of Temperature Variables



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

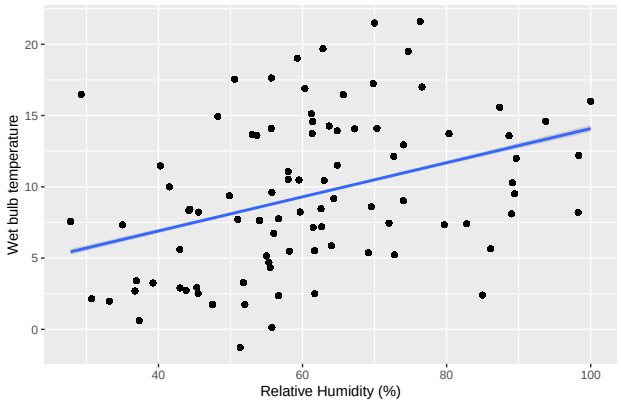
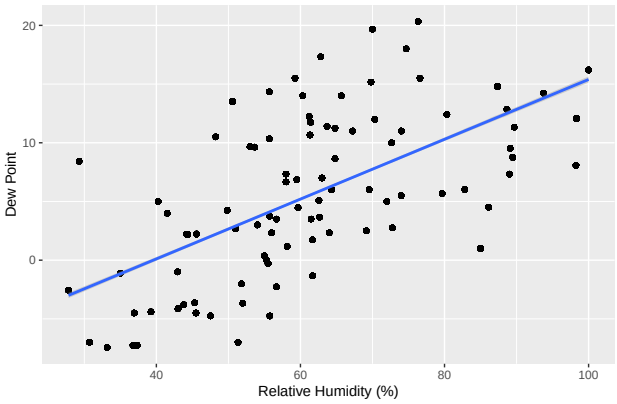


Figure 3b. Relationship between Relative Humidity and Dew Point



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

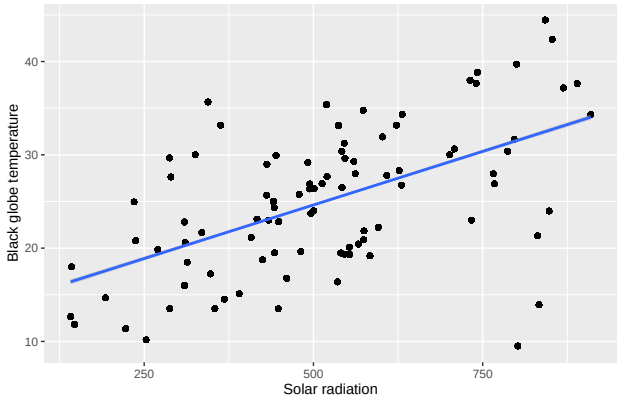


Figure 5 shows the changes in marathon performance with age, where the black line represents the linear

fit and the red dashed line represents the mean. We found that as age increases, performance deteriorates and shows a U-shaped trend. Moreover, this trend does not seem to have a significant gender difference. Specifically, their performance peaks around the age of 25-50 and then declines with increasing age. If we carefully observe the tails, we find that men seem to rise faster, which means that as age increases, men's performance declines faster than women's.

Table 3 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 3: 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. Specifically, between the ages of 70-79 for

males and 80-91 for females, the performance is actually improved with increasing humidity. This pattern is unique and displays gender differences, indicating the unique adaptation or tolerance of older athletes to humidity.

Figure 6e shows the impact of air quality index on marathon performance for different age groups. Analysis shows that for participants aged 14 to 69, regardless of gender, their performance deteriorates as AQI increases. However, for participants aged 70 to 91, as AQI increases, their performance also improves. This counterintuitive trend may be attributed to insufficient sample size among the elderly population and a lack of sufficient high AQI samples. This may introduce bias into the research results.

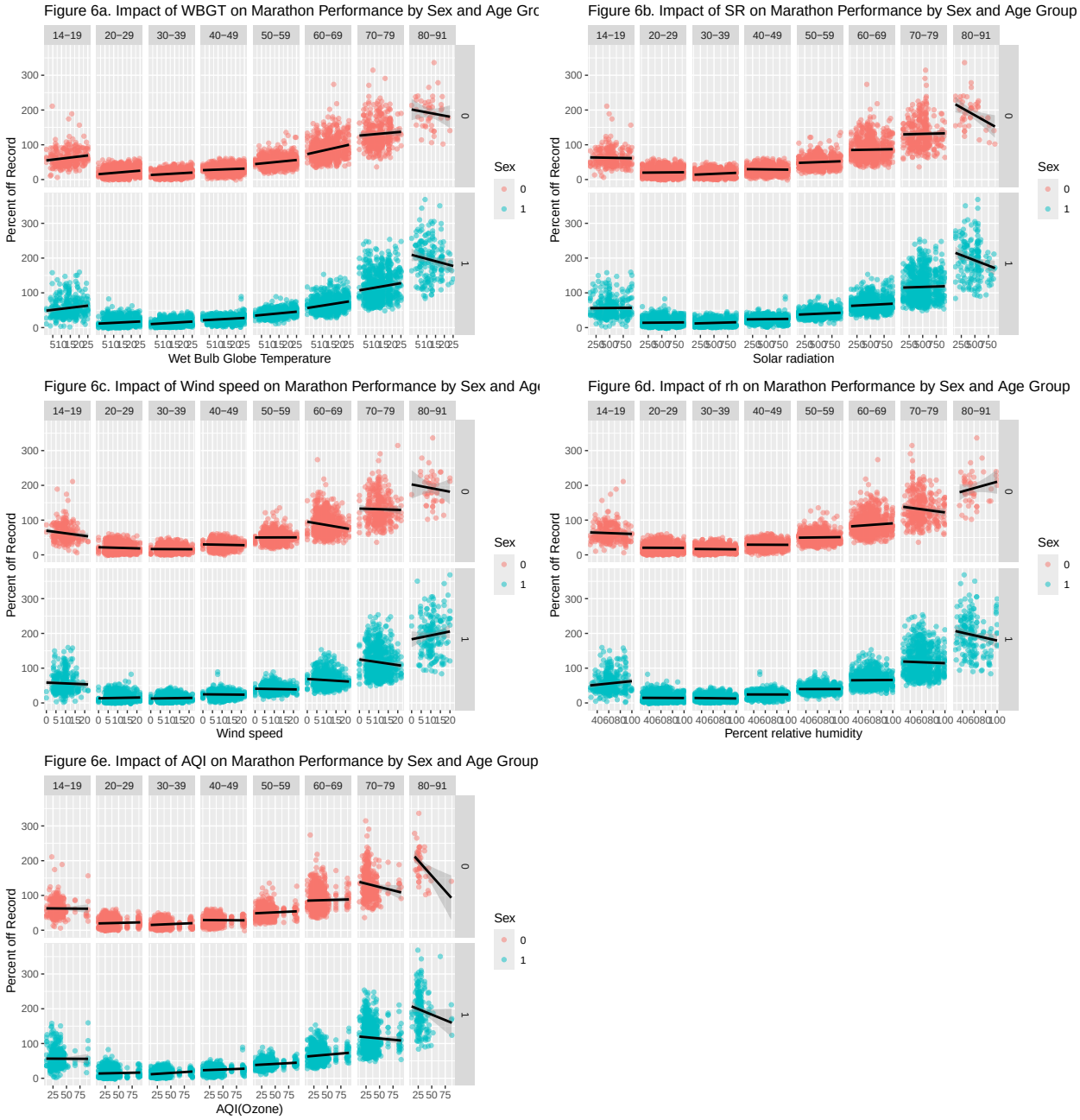


Table 4 provides a summary analysis of the impact of different WBGT levels on performance across various age groups and genders. As observed in Figure 6a, performance generally deteriorates with rising temperatures for both males and females. Notable exceptions include males aged 14-19, who perform better under red conditions than yellow, and all participants aged 80-91, whose performance under yellow and red conditions

Table 4: 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.

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, AQI, and the two-way interaction terms of Age, Sex, and these variables, as well as Age^2. Due to the 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 5 shows that environmental conditions significantly influence marathon performance, and these effects vary based on gender and age. Specifically, high humidity (`rh`), solar radiation (`SR`), wind speed (`Wind`), temperature (high `WBGT`), and poor air quality (`AQI`) are associated with differences in performance, while higher dew points (`DP`) appear to improve performance. Moreover, the impact of environmental conditions appears to diminish with age, and gender differences suggest that males may be less resilient to certain environmental conditions (except `DP`).

Specifically, the coefficient of `rh` is -0.0029, suggesting that for each unit increase in relative humidity, the log-transformed marathon performance (`Percent_off_log`) decreases by approximately 0.0029 units. Since lower `Percent_off_log` values indicate better performance, this implies that higher humidity correlates with better performance. The coefficient of `SR` is 0.00014, indicating that each unit increase in solar radiation slightly worsens performance, as `Percent_off_log` increases by 0.00014 units. The coefficient of `DP` is 0.0193, showing that each unit increase in dew point improves performance, as `Percent_off_log` decreases by approximately 0.0193 units. The coefficient of `Wind` is -0.0054, indicating that each unit increase in wind speed improves performance, as `Percent_off_log` decreases by approximately 0.0054 units. The coefficient of `WBGT` is 0.0140, demonstrating that each unit increase in WBGT corresponds to a deterioration in performance, as `Percent_off_log` increases by about 0.0140 units. Lastly, the coefficient of `AQI` is -0.0024, indicating that worse air quality (higher AQI) corresponds to better performance, though this result may be counterintuitive and warrant further investigation.

The coefficient of `Age:Sex1` is 0.0027, suggesting that as age increases, the rate of performance improvement associated with male gender diminishes compared to females. The coefficient for `Sex1:rh` is 0.00163, indicating that, relative to females, males benefit less from increased humidity. The coefficient for `Sex1:DP` is -0.00768, meaning that higher dew points benefit females more significantly than males. The coefficient

Table 5: Regression Results Summary

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.1835	0.0102	213.1235	0e+00
Age	0.0380	0.0002	194.1409	0e+00
Sex1	-0.5343	0.0042	-128.1449	0e+00
rh	-0.0029	0.0001	-30.5009	0e+00
SR	0.0001	0.0000	37.5035	0e+00
DP	0.0193	0.0005	40.2748	0e+00
Wind	-0.0054	0.0002	-28.3826	0e+00
WBGT	0.0140	0.0005	28.5703	0e+00
AQI	-0.0024	0.0000	-56.8592	0e+00
Age:Sex1	0.0027	0.0000	136.8846	0e+00
Sex1:rh	0.0016	0.0000	41.7106	0e+00
Sex1:SR	0.0000	0.0000	3.9601	1e-04
Sex1:DP	-0.0077	0.0002	-40.3390	0e+00
Sex1:Wind	0.0041	0.0001	56.8465	0e+00
Sex1:WBGT	0.0100	0.0002	49.7723	0e+00
Sex1:AQI	0.0006	0.0000	31.5485	0e+00
Age:rh	0.0000	0.0000	-9.3401	0e+00
Age:SR	0.0000	0.0000	-68.3868	0e+00
Age:DP	-0.0002	0.0000	-15.3929	0e+00
Age:Wind	0.0001	0.0000	26.6117	0e+00
Age:WBGT	-0.0002	0.0000	-23.2307	0e+00
Age:AQI	0.0000	0.0000	28.8746	0e+00

Table 6: Comparison of Model AIC: With and Without Age2

Model	AIC
Without Age^2	-1371.5454
With Age^2	141.2683

for `Age:rh` is -0.0000173, indicating that the positive impact of humidity on performance becomes less pronounced with age. The coefficient for `Age:SR` is -0.0000049, suggesting that older individuals are slightly more resilient to solar radiation. The coefficient for `Age:DP` is -0.000158, indicating that the positive effects of higher dew points diminish with age. The coefficient for `Age:AQI` is 0.0000211, showing that the unexpected performance benefits of higher AQI decrease slightly with age.

Table 6 shows the model without Age^2 is preferable based on the AIC values, as it achieves a much better balance between fit and complexity. Adding Age^2 does not improve the model sufficiently to justify the increased complexity.

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 Conditions by Sex

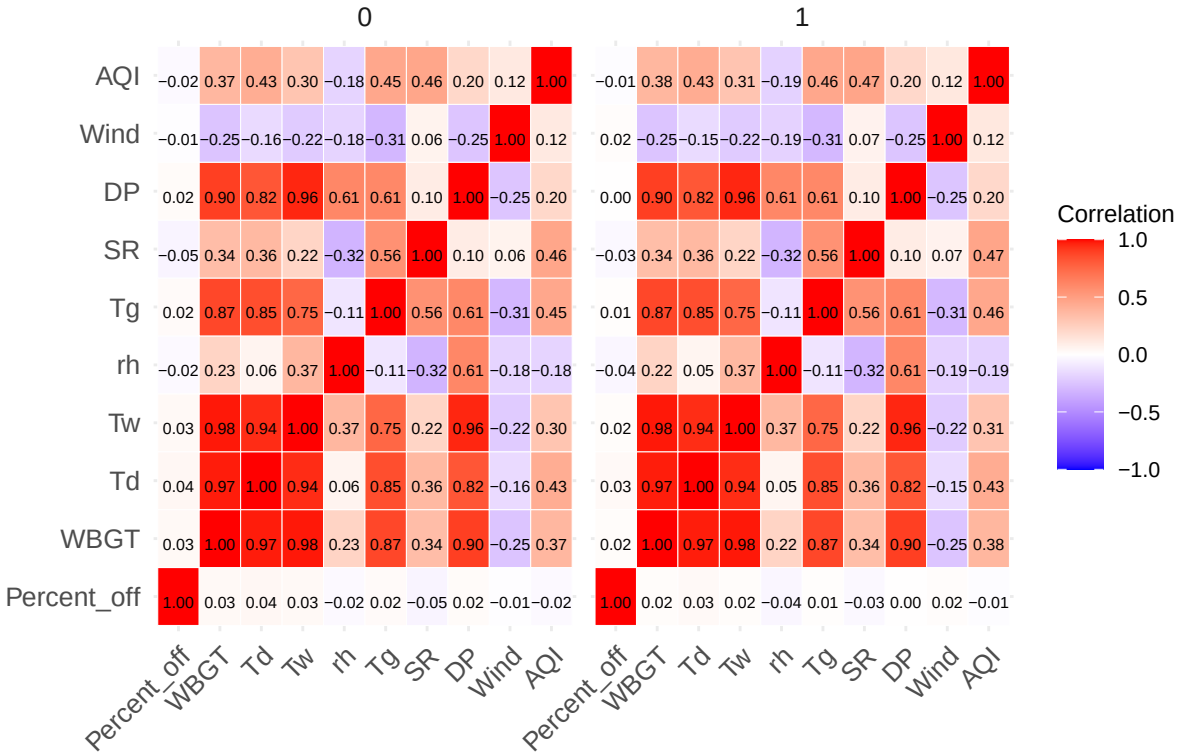
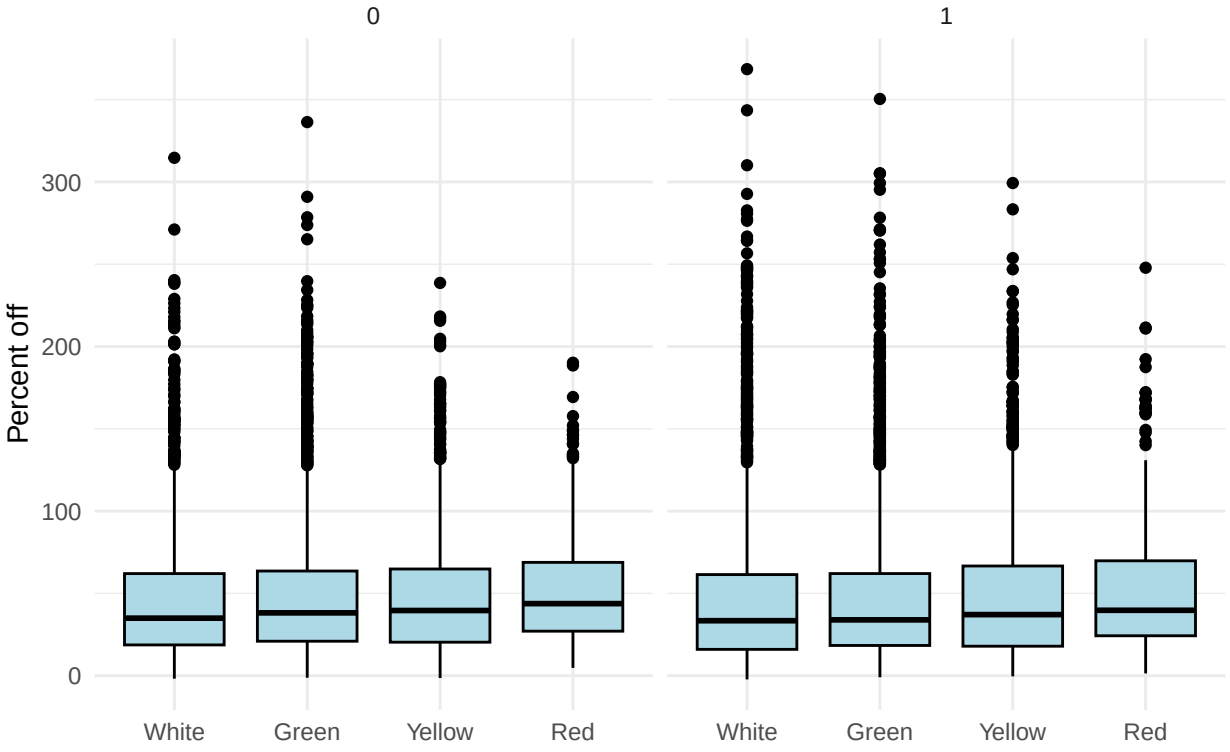


Table 7: Stepwise Regression Results: Impact of Removing Variables on Model Fit

Variable	RSS	AIC
none	11060	10.5
Tg	11163	111.7
Td	11505	445.5
DP	11714	644.7
Tw	11849	772.1
rh	12795	1622.5
AQI	13196	1964.2
Wind	18836	5904.6
Flag	25720	9349.7
SR	27979	10285.9

Figure 8. Flag conditions and Percent off



We utilized stepwise regression to analyze which environmental conditions have the greatest impact on marathon performance. Table 6 indicates that removing SR (Solar Radiation) resulted in an increase in the residual sum of squares (RSS) to 27979 and an increase in the AIC to 10285.9. This suggests that SR is one of the most influential variables contributing to the model. Additionally, Flag (indicating environmental risk level based on WBGT), AQI, 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 7 shows that rh and DP (Dew Point) are the most impactful variables on marathon performance, as they

Table 8: Coefficient Analysis without WBGT and Flag

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.6363	0.0001	24909.0317	0
Tw	-0.0126	0.0016	-7.7099	0
Tg	0.0134	0.0007	19.8897	0
Td	-0.0157	0.0018	-8.8966	0
rh	-0.0991	0.0011	-93.8056	0
SR	-0.0507	0.0003	-146.9659	0
DP	0.1208	0.0013	93.9466	0
Wind	0.0188	0.0002	90.7546	0
AQI	-0.0111	0.0001	-106.3598	0

Table 9: Coefficient Analysis with WBGT

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.5668	0.0008	4541.9415	0
FlagGreen	0.0783	0.0012	65.2323	0
FlagYellow	0.1219	0.0017	71.6713	0
FlagRed	0.2691	0.0022	124.8870	0
rh	-0.0826	0.0010	-86.6097	0
SR	-0.0466	0.0003	-145.0034	0
DP	0.0803	0.0022	36.3455	0
Wind	0.0237	0.0002	97.7987	0
WBGT	-0.0411	0.0022	-18.6933	0
AQI	-0.0135	0.0002	-72.8365	0

have the largest coefficients and the highest t-values, indicating their strong influence within the model. SR follows closely, with both its t-value and coefficient underscoring its importance.

Table 8 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 effects in the model. Under certain environmental conditions, an increase in SR may help improve performance by optimizing athletes' heat dissipation mechanisms. The higher the DP, the worse the athlete's performance.

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.

6. Mccrowey C, Sharac T, Mangus N, Jager D, Brown R, Garver D, Wells B, Brittingham H (2023). A R Interface to the US EPA Air Quality System Data Mart API. https://cran.r-project.org/package=RAQSAPI.

Code Appendix:

```
## load packages
knitr::opts_chunk$set(echo = FALSE)
library(tidyverse)
library(naniar)
library(dplyr)
library(knitr)
library(tidyr)
library(reshape2)
library(kableExtra)
library(ggplot2)
library(gtsummary)
# read data
data <- readRDS('/Users/fusei/Desktop/24FALL/PHP2550/Project1/project1.cleaned_data.rds')

## use complete data
data <- data[complete.cases(data),]
#dim(data) ## 11073 x 14
## Correct Relative humidity
p1 <-
  ggplot(data, aes(x = rh)) +
  geom_histogram(binwidth = 1, color = "black", fill = "blue") +
  labs(title = "Figure 1a: Histogram of relative humidity", x = "rh", y = "Frequency") +
  theme_minimal()

## rescale rh <= 1
data$rh <- ifelse(data$rh <= 1,data$rh*100, data$rh)

## make a histogram
p2 <-
  ggplot(data, aes(x = rh)) +
  geom_histogram(binwidth = 1, color = "black", fill = "blue") +
  labs(title = "Figure 1b: Histogram of corrected relative humidity", x = "rh", y = "Frequency") +
  theme_minimal()

## show
print(p1)
print(p2)
# Convert the Flag column into a factor with a specified order of levels
data_table2 <- data %>%
  mutate(
      # Relevel Flag as a factor with the specified order
    Flag = factor(Flag, levels = c("White", "Green", "Yellow", "Red"))
  )


# Create a summary table for environmental data grouped by "Flag"
table1 <- tbl_summary(
  # Select relevant columns for the summary table
  data_table2 %>% select(Flag, Td, Tw, rh, Tg, SR, DP, Wind, WBGT),
  # Group data by "Flag"
  by = "Flag",
```

```r
  # Specify the type of variables to summarize
  type = list(
    c(Td, Tw, rh, Tg, SR, DP, Wind, WBGT) ~ "continuous2"
  ),

  # Define the summary statistics for continuous variables
  statistic = list(
    all_continuous() ~ c("{mean}", "{min}, {max}")
  ),
  # Set the number of digits for continuous variables
  digits = list(
    all_continuous() ~ c(2, 2)
  ),
  # Assign custom labels for each variable
  label = list(
    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
# Convert the summary table to a LaTeX table with additional formatting
   as_kable_extra(booktabs = TRUE,
                  caption = "Summary Statistics of Environmental Information by Flag",
                 longtable = TRUE, linesep = "") %>%
    # Style the LaTeX table for better appearance
  kableExtra::kable_styling(font_size = 8,
                            latex_options = c("repeat_header", "HOLD_position"))
# show
table1
# Preprocess the data by relabeling factors for Race, Sex, and Flag
data_table1 <- data %>%
  mutate(
    # Convert Race column to a factor with custom labels
    Race = factor(Race, levels = 0:4,
                  labels = c("Boston", "Chicago", "NYC", "Twin Cities", "Grandma's")),

    # Convert Sex column to a factor with custom labels
    Sex = factor(Sex, levels = c(0, 1), labels = c("Female", "Male")),

    # Convert Flag column to a factor with specified levels
    Flag = factor(Flag, levels = c("White", "Green", "Yellow", "Red"))
  )

# Create a summary table for personal information grouped by "Flag"
table_personal <- tbl_summary(
  # Select relevant columns for the summary table
  data_table1 %>% select(Flag, Race, Sex, Age, Percent_off),
```

```r
  # Group data by "Flag"
  by = "Flag",

  # Specify the type of variables to summarize
  type = list(
    Age ~ "continuous2",  # Age is treated as a continuous variable
    Percent_off ~ "continuous2"  # Percent_off is treated as a continuous variable
  ),

  # Define the summary statistics for continuous variables
  statistic = list(
    all_continuous() ~ c("{mean}", "{min}, {max}")  # Show mean, min, and max
  ),

  # Set the number of decimal places for continuous variables
  digits = list(
    all_continuous() ~ c(2, 2)  # Round to 2 decimal places
  ),

  # Assign custom labels for each variable in the summary table
  label = list(
    Race ~ "Race",
    Sex ~ "Sex",
    Age ~ "Age",
    Percent_off ~ "Percent off Current Record"
  )
) %>%
  # Convert the summary table to a LaTeX table with additional formatting
  as_kable_extra(
    booktabs = TRUE,  # Use booktabs for a professional table style
    caption = "Summary Statistics of Personal Information by Flag",  # Add a table caption
    longtable = TRUE,  # Allow the table to span multiple pages in LaTeX
    linesep = ""  # Remove extra line separators
  ) %>%
  # Style the LaTeX table for better appearance
  kableExtra::kable_styling(
    font_size = 8,  # Set font size to 8 for readability
    latex_options = c("repeat_header", "HOLD_position")  # Repeat headers and hold position
  )

# Display the final table
table_personal
## Single Variable: Age
# Plot the distribution of Age as a histogram
ggplot(data, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +  # Histogram with blue fill and black
  ggtitle("Figure 2a. Distribution of Age") +  # Add a title to the plot
  xlab("Age") +  # Label for the x-axis
  ylab("Frequency")  # Label for the y-axis

## Age by Gender
# Plot the distribution of Age by gender using faceted histograms
ggplot(data, aes(x = Age)) +
```

```r
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +  # Histogram with blue fill and black
  ggtitle("Figure 2b. Distribution of Age") +  # Add a title to the plot
  facet_wrap(~Sex) +  # Create separate plots for each gender
  xlab("Age") +  # Label for the x-axis
  ylab("Frequency")  # Label for the y-axis

## Performance: Percent Off
# Create a boxplot to show the distribution of Percent_off
ggplot(data, aes(x = Percent_off)) +
  geom_boxplot(fill = "skyblue") +  # Boxplot with sky blue fill
  ggtitle("Figure 2c. Distribution of Percent off") +  # Add a title to the plot
  xlab("") +  # Empty label for the x-axis
  ylab("Percent off %")  # Label for the y-axis

## TW, DP, and WBGT
# Create a histogram to show the distribution of temperature-related variables
p <- ggplot(data, aes(x = value)) +
  geom_histogram(binwidth = 1, fill = "skyblue", color = "black") +  # Histogram with sky blue fill and
  theme_minimal() +  # Apply a minimal theme
  labs(x = "Temperature ", y = "Frequency") +  # Add labels to the axes
  facet_wrap(~variable, scales = "free")  # Facet the plot by variable, allowing free scales

# Reshape data to long format for boxplot visualization
data_long <- data %>%
  select(Tw, DP, WBGT) %>%  # Select columns for Tw, DP, and WBGT
  pivot_longer(cols = everything(), names_to = "variable", values_to = "value")  # Convert to long form

# Create boxplots for Tw, DP, and WBGT
ggplot(data_long, aes(x = "", y = value)) +
  geom_boxplot(fill = "lightblue", color = "black") +  # Boxplot with light blue fill and black borders
  facet_wrap(~variable, scales = "free_y") +  # Facet the plot by variable, allowing free y-axis scales
  labs(
    title = "Figure 2d. Distribution of Temperature Variables",  # Add a title to the plot
    x = "",  # Empty label for the x-axis
    y = "Temperature "  # Label for the y-axis
  ) +
  theme_minimal()  # Apply a minimal theme

# Plot the distribution of 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
# Plot the distribution of 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 Humidity

# Plot the distribution of SR and Tg
ggplot(data, aes(x = SR, y = Tg)) +
  geom_point() +
  geom_smooth(method = "lm") +
```

```r
  labs(title = "Figure 4. Relationship between Solar radiation and Wet Bulb", x = "Solar radiation", y =


# Calculate Mean Percent off
mean_lines <- data %>%
  group_by(Age, Sex) %>%
  summarize(mean_percent_off = mean(Percent_off, na.rm = TRUE), .groups = "drop")

# Age vs. Performance by sex
ggplot(data, aes(x = Age, y = Percent_off, color = Sex)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", aes(linetype = "Fitted Line"), se = TRUE, color = "black") +
  geom_line(data = mean_lines, aes(x = Age, y = mean_percent_off, linetype = "Mean Line"),
            size = 1, color = "red") +
  labs(
    title = "Figure 5. Age vs. Marathon Performance by Sex",
    x = "Age",
    y = "Percent off Current Record",
    linetype = "Line Type"
  ) +
  facet_wrap(~Sex) +
  scale_linetype_manual(values = c("Fitted Line" = "solid", "Mean Line" = "dashed")) +
  theme_minimal()


# Add a new column 'Age_Group' by categorizing the 'Age' variable into predefined age ranges
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"  # For any values outside the defined ranges
    )
  )

# Calculate summary statistics (mean and standard deviation) of 'Percent_off' by 'Sex' and 'Age_Group'
summary_table <- data_table2 %>%
  group_by(Sex, Age_Group) %>%  # Group data by 'Sex' and 'Age_Group'
  summarise(
    Mean_Performance = mean(Percent_off, na.rm = TRUE),  # Calculate mean, ignoring missing values
    SD_Performance = sd(Percent_off, na.rm = TRUE),  # Calculate standard deviation, ignoring missing v
    .groups = 'drop'  # Ungroup after summarisation
  ) %>%
  # Combine mean and SD into a single string with "±" formatting
  mutate(Performance = paste(round(Mean_Performance, 2), "±", round(SD_Performance, 2)))

# Keep only the 'Sex', 'Age_Group', and combined 'Performance' columns
```

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

# Reshape the summary table into a wide format
summary_wide <- summary_table %>%
  pivot_wider(
    names_from = Age_Group,  # Create columns for each 'Age_Group'
    values_from = Performance  # Fill the new columns with 'Performance' values
  )

# Generate a LaTeX table from the wide-format summary table
kable_output <- summary_wide %>%
  kable(
    "latex",  # Specify LaTeX output format
    booktabs = TRUE,  # Use booktabs for professional table styling
    longtable = TRUE,  # Allow the table to span multiple pages
    caption = "Age Group vs. Average Performance by Sex"  # Add a caption for the table
  ) %>%
  # Apply additional styling to the LaTeX table
  kableExtra::kable_styling(
    font_size = 8,  # Set font size to 8 for better readability
    latex_options = c("repeat_header", "HOLD_position")  # Repeat headers on new pages and hold positio
  )

# Display the final table
kable_output
# Add a new column 'Age_Group' by categorizing the 'Age' variable into predefined age ranges
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"
  ))
### Plot the distribution of  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")
### Plot the distribution of SR vs. Performance:
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")
### Plot the distribution of Wind vs. Performance:
```

```r
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")

### Plot of 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")

### Plot of AQI vs. Performance:
ggplot(data_figure6, aes(x = AQI, 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 6e. Impact of AQI on Marathon Performance by Sex and Age Group",
       x = "AQI(Ozone)", y = "Percent off Record")

# Step 1: Data Preparation
# Categorize 'Age' into predefined groups and calculate performance statistics by Flag, Sex, and Age_Gr
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'
  ) %>%
    # Combine mean and standard deviation into a single summary column
  mutate(Percent_off_Summary = paste(round(Mean_Percent_off, 2), "±", round(SD_Percent_off, 2)))
# Step 2: Create unique identifiers for Age_Group and Sex
data_table3 <- data_table3 %>%
  mutate(Age_Sex = str_c("Age_", Age_Group, "_Sex_", Sex))
# Step 3: Reshape the data to wide format
data_wide <- data_table3 %>%
  pivot_wider(names_from = Age_Sex, values_from = Percent_off_Summary)
# Step 4: Initialize a data frame with unique 'Flag' values
data_table3 <- data.frame(Flag = unique(data_wide$Flag))
```

```r
# Step 5: Populate the table with relevant columns for each Age_Sex combination
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]]
}
# Step 6: Reorder and split the table into two parts for display
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_
# Split the data into two tables for easier visualization
data_table3_1 <-data_table3[,1:9]
data_table3_2 <-data_table3[,c(1,10:17)]
# Step 7: Generate LaTeX Table for the First Part
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
                              ))
# Step 7: Generate LaTeX Table for the Second Part
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))

#Display
kable_output3_1
kable_output3_2

### model without age^2
# Add a new column: log-transformed Percent_off to address skewness and stabilize variance
data$Percent_off_log = log(data$Percent_off + 4)
# Fit a linear model (without Age^2) to predict log-transformed Percent_off
model <- lm(Percent_off_log ~ Age * Sex +
               rh + SR + DP + Wind + WBGT + AQI +
               Sex * rh + Sex * SR + Sex * DP + Sex * Wind + Sex * WBGT + Sex * AQI +
               Age * rh + Age * SR + Age * DP + Age * Wind + Age * WBGT + Age * AQI  ,
            data = data)
# Calculate weights based on the inverse of squared residuals for weighted least squares (WLS)
weights <- 1 / residuals(model)^2
# Fit a weighted least squares model with the calculated weights
wls_model <- lm(Percent_off_log ~ Age * Sex +
                   rh + SR + DP + Wind + WBGT + AQI +
                   Sex * rh + Sex * SR + Sex * DP + Sex * Wind + Sex * WBGT + Sex * AQI +
                   Age * rh + Age * SR + Age * DP + Age * Wind + Age * WBGT + Age * AQI ,
                data = data, weights = weights)
# Extract the regression results summary as a data frame
```

```r
coefficients_df <- as.data.frame(coef(summary(wls_model)))
# Create a LaTeX table to display the regression coefficients
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
## model with age^2
# Fit a linear model (with Age^2) to predict log-transformed Percent_off
model2 <- lm(Percent_off_log ~ Age * Sex +
                rh + SR + DP + Wind + WBGT + AQI +
                Sex * rh + Sex * SR + Sex * DP + Sex * Wind + Sex * WBGT + Sex * AQI +
                Age * rh + Age * SR + Age * DP + Age * Wind + Age * WBGT + Age * AQI +
                I(Age^2) ,
             data = data)
# Calculate weights based on the inverse of squared residuals for weighted least squares (WLS)
weights2 <- 1 / residuals(model2)^2
# Fit a weighted least squares model with the calculated weights
wls_model2 <- lm(Percent_off_log ~ Age * Sex +
                    rh + SR + DP + Wind + WBGT + AQI +
                    Sex * rh + Sex * SR + Sex * DP + Sex * Wind + Sex * WBGT + Sex * AQI +
                    Age * rh + Age * SR + Age * DP + Age * Wind + Age * WBGT + Age * AQI + I(Age^2),
                 data = data, weights = weights2)
# Calculate AIC
aic_values <- data.frame(
  Model = c("Without Age^2", "With Age^2"),
  AIC = c(AIC(wls_model), AIC(wls_model2))
)
# Create a LaTeX table to display the regression coefficients
kable_table <- kable(
  aic_values,
  format = "latex",
  col.names = c("Model", "AIC"),
  caption = "Comparison of Model AIC: With and Without Age2"
) %>%
  kable_styling(bootstrap_options = c("striped", "hover"), full_width = FALSE, position = "center", fon

kable_table
## continuous variabls by sex
# Compute correlation matrices for each sex
numeric_columns <- c("Percent_off", "WBGT", "Td", "Tw", "rh", "Tg", "SR", "DP", "Wind","AQI")
# Group by sex
cor_by_sex <- data %>%
  group_by(Sex) %>%
  summarise(cor_matrix = list(cor(across(all_of(numeric_columns)), use = "complete.obs")))

# Melt the correlation matrices
melted_cor_by_sex <- cor_by_sex %>%
  mutate(melted = map(cor_matrix, ~ melt(.x, varnames = c("Var1", "Var2")))) %>%
  select(Sex, melted) %>%
```

```r
  unnest(melted)

# Plot faceted heatmaps
ggplot(melted_cor_by_sex, aes(Var1, Var2, fill = value)) +
  geom_tile(color = "white") +
  geom_text(aes(label = sprintf("%.2f", value)), vjust = 1, size =2) +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1, 1), space
  theme_minimal() +
  labs(x = "", y = "", title = "Figure 7. Heatmap of Correlation Between Percent off and Weather Conditi
  facet_wrap(~Sex) +
    theme(
    axis.text.x = element_text(size = 10, angle = 45, hjust = 1),
    axis.text.y = element_text(size = 10),
    strip.text = element_text(size = 10),
    legend.text = element_text(size = 8),
    legend.title = element_text(size = 9),
    plot.title = element_text(size = 8)
  )



## Display 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") +
  facet_wrap(~Sex) +
  theme_minimal()
# log-transformed Percent_off
data$Percent_off_log = log(data$Percent_off + 4)
# Fit a linear model to predict log-transformed Percent_off
model <- lm(Percent_off_log ~ Tw + Tg + Td + Flag +
               rh + SR + DP + Wind + WBGT + AQI ,
            data = data)
# Calculate weights based on the inverse of squared residuals for weighted least squares (WLS)
weights <- 1 / residuals(model)^2
# Fit a weighted least squares model with the calculated weights
wls_model <- lm(Percent_off_log ~ Tw + Tg + Td + Flag +
                    rh + SR + DP + Wind + WBGT + AQI,
                data = data, weights = weights)
# Backward Model Selection
backward_model <- step(wls_model, direction = "backward",trace = 0)

## Show stepwise output
stepwise_results <- data.frame(
  Variable = c("none","Tg", "Td", "DP", "Tw", "rh", "AQI",   "Wind", "Flag", "SR"),
  RSS = c(11060, 11163, 11505, 11714, 11849, 12795, 13196, 18836,  25720, 27979),
  AIC = c(10.5, 111.7, 445.5, 644.7,772.1, 1622.5, 1964.2,  5904.6,  9349.7, 10285.9)
)

# Create a LaTeX table to display the regression coefficients
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
# Log-transform the Percent_off variable to address skewness and stabilize variance
data$Percent_off_log = log(data$Percent_off + 4)
# Standardize the environmental variables for comparability
data_scaled <- data.frame(scale(data[, c("Tg", "rh", "Tw", "DP", "Wind", "Td", "SR", "WBGT", "AQI")]))
data_scaled$Percent_off_log <- data$Percent_off_log
data_scaled$Flag <- data$Flag
## Model fitting without WBGT
# Fit an ordinary least squares (OLS) linear model without the WBGT variable
model <- lm(data_scaled$Percent_off_log ~ Tw + Tg + Td +
              rh + SR + DP + Wind + AQI ,
            data = data_scaled)
# Calculate weights for weighted least squares (WLS) based on the inverse of residual variances
weights <- 1 / residuals(model)^2
# Fit with weights
wls_model <- lm(data_scaled$Percent_off_log ~ Tw + Tg + Td +
                  rh + SR + DP + Wind  + AQI ,
                data = data_scaled, weights = weights)
# Extract coefficients
coefficients_df <- as.data.frame(coef(summary(wls_model)))
# Creat table
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
# Fit an ordinary least squares (OLS) linear model witt the WBGT variable
modelw <- lm(data_scaled$Percent_off_log ~  Flag +
               rh + SR + DP + Wind + WBGT + AQI,
             data = data_scaled)
# Calculate weights
weights <- 1 / residuals(modelw)^2
# Fit model
wls_modelw <- lm(data_scaled$Percent_off_log ~  Flag +
                   rh + SR + DP + Wind + WBGT + AQI,
                 data = data_scaled, weights = weights)
# Extract coefficients
coefficients_df <- as.data.frame(coef(summary(wls_modelw)))
# Create Table
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
```