Exploring the Role of Weather and Environmental Variables on Marathon Performance Across Age and Gender

Yingxi Kong

Abstract

Environmental condition characteristics are suspected to be factors of runners' performance during Marathons. Using data collected over years across five major U.S. marathons Boston, Chicago, New York City, Twin Cities, and Grandma's, this study aims to investigate the effect of environmental characteristics, including key variables like Wet Bulb Globe Temperature (WBGT), solar radiation (SRWm2), and relative humidity (rh), on runner's performance which is measured as the percentage off the best course record (%CR), and how their effect varies by gender and age. Performing comprehensive exploratory data analysis and statistical modeling, our results indicates that age has a significant effect on runners' performance and male runners show more sensitivity of age variation. In addition, environmental conditions like WBGT, solar radiation, and relative humidity do slightly affect runner's performance. Older people in upper-mid or higher age group and younger people aging from 14-25 years old show more sensitivity as environmental conditions vary. Female runners exhibits more stable performance under varying environmental conditions compared to male runners. However, our analysis is still limited by a smaller number of observations in the highest age group (70+) and younger age group (14-25), suggesting the need for further research with a larger sample size. Participants of Boston Marathon may represent a more experienced sample, potentially limiting the generalizability of less experienced runners in our study.

Introduction

Marathon running, a long-distance race covering 42.195 kilometers (26.2 miles), has been known for its physical and mental challenges. Over the years, it has attracted an increasing number of excellent runners from around the world, leading to interests in how physiological and environmental conditions influence runner's performance. Understanding these factors would help runners to better understand their performance. Also, it would provide more information to the runners to optimize their training plans and race-day strategies, leading to improvement of their performance.

This report is a collaboration with Dr. Brett Romano Ely and Dr. Matthew Ely from the Department of Health Sciences at Providence College, which explores how environmental conditions, age, and sex would influence runner's performance in this long-distance race. Their prior research found that warmer temperature leads to decline in performance in marathon races, and this decline in endurance performance varies significantly between females and males. Moreover, older adults face more thermoregulatory challenges during exercise, which further exacerbate performance declines under warmer temperature. This exploratory analysis study aims to build on previous findings, providing deeper insight on the following aims:

- Aim 1: Examine effects of increasing age on marathon performance in men and women.l
- Aim 2: Explore the impact of environmental conditions on marathon performance, and whether the impact differs across age and gender.
- Aim 3: Identify the weather parameters (WBGT, Flag conditions, temperature, etc) that have the largest impact on marathon performance.

Data Collection

This project combines information from four datasets, including information of participants' performance, air quality, and the corresponding information for each race. The primary dataset in this project consists of 11,564 observations with 14 variables, including information on participants' characteristics and environmental condition characteristics collected from five major marathon races over a period of 15-20 years in the U.S.: the Boston Marathon, Chicago Marathon, New York City Marathon, Twin Cities Marathon (Minneapolis, MN), and Grandma's Marathon (Duluth, MN). Our response variable is `CR_pct`, representing the percent off the current course record (%CR). Key covariates include race, sex, temperature (`Tdc`, `Twc`, `Tgc`, etc.), relative humidity (`rh`), solar radiation (`SRWm2`), dew point (`DP`), wind speed (`Wind`), and Wet Bulb Globe Temperature (`WBGT`).

The air quality data from the `RAQSAPI` package consists of 10,451 observations and 11 variables, containing detailed information on Air Quality Index (AQI) for each race day over years. The course record dataset has 194 rows and 4 columns, providing information on the best course record for each marathon race by gender. The date dataset includes information of specific date on which each races were held with 98 observations and 3 variables.

Preprocessing

The main data set has 491 observations with missing values, and these missingness appears mainly among environmental conditions covariates from races held in 2011 and 2012. These missing observations were neither excluded from the analysis nor imputed to retain as much data as possible for exploration.

During the preprocessing steps, we managed the column names for easier understanding, modified columns with coding issues, and ensured the correct data types for various columns. In addition, we merged the three additional datasets with our primary data to add in more information we are interested in.

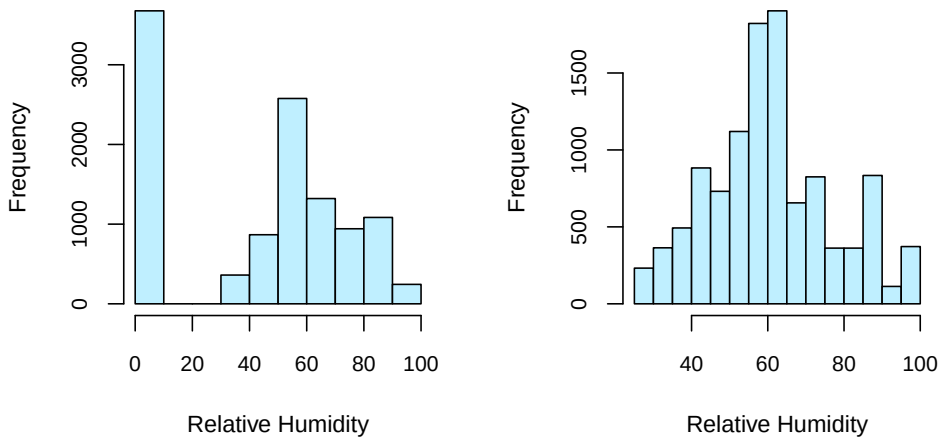
Moreover, we mutated a new categorical variable `age_group` for further aging analysis. Based on the age distribution in the dataset, runners were classified into five aging groups: Younger, Lower-Mid Age, Mid Age, Upper-Mid Age, and Highest Age with cut-off points presented in `Table 1`. These cut-off points are decided by considering both the natural breaks in the data and common practices in marathon age classifications. Instead of using the traditional 10-year groups, we set 15-year groups to better reflect the age distribution of participants in our dataset.

Table 1: Age Group Classifications for Marathon Participants

Group	Age Range
Younger	Below 25
Lower-Mid Age	25-39
Mid Age	40-54
Upper-Mid Age	55-69
Highest Age	70+

Additionally, during our data investigation, we found there are nearly half of the observations in the relative humidity column (`rh`) have near-zero values, which is not reasonable. As shown in `Figure 1`, there might be issues with data collection and recoding process. To fix this, We multiplied those near-zero values by 100 to ensure they are within a realistic range.

Figure 1: Distribution of Relative Humidity Before and After Correction



Summary Statistics

To perform data investigation, we first create a summary table of all environmental conditions experienced during the five major marathons in `Table 2`. The Boston Marathon was the only race with a completed record of weather conditions. Observing the pattern of weather conditions, the Grandma's Marathon exhibited the warmest temperature where they had the highest average dry bulb temperature ($18.9°C$), wet bulb temperature ($14.9°C$), and black global temperature($32°C$), reflecting a higher risk of heat illness. Moreover, the Grandma's Marathon had 47% of time falling within the WGBT range of $18 - 23$ $°C$. In contrast, the Boston and New York City Marathons had the coolest temperature with more races falling within the WGBT range of < 10 $°C$. Similarly, these two Marathons presented lowest average dry bulb, wet bulb, and black global temperatures. The Chicago and Twin Cities Marathons had relatively moderate temperature with more races falling within the WGBT range of $10 - 18$ $°C$. Wind speeds were higher in Boston and New York City and remain consistent around 9 km/h for other races. Boston Marathon exhibited highest air quality index (`aqi`) while the New York City Marathon exhibited the lowest. Moreover, the Grandma's Marathon had the highest level of relative humidity and New York City Marathon had the lowest. Based on this summary statistics, we observed the significant variation in environmental conditions across different location which might have a substantial impact on the runners' performance.

Table 2: Summary Table of Weather Parameters

Characteristic	**Race**				
	Boston Marathon, N = 18	**Chicago Marathon, N = 21**	**Grandma's Marathon, N = 17**	**New York City Marathon, N = 23**	**Twin Cities Marathon, N = 17**
flag					
Missing	0 (0%)	1 (4.8%)	1 (5.9%)	1 (4.3%)	1 (5.9%)
WBGT < 10C	9 (50%)	6 (29%)	0 (0%)	11 (48%)	5 (29%)
WBGT > 18-23C	1 (5.6%)	1 (4.8%)	8 (47%)	4 (17%)	3 (18%)
WBGT > 23-28C	1 (5.6%)	1 (4.8%)	2 (12%)	0 (0%)	1 (5.9%)

Table 2: Summary Table of Weather Parameters *(continued)*

Characteristic	Race				
	Boston Marathon, N = 18	**Chicago Marathon, N = 21**	**Grandma's Marathon, N = 17**	**New York City Marathon, N = 23**	**Twin Cities Marathon, N = 17**
WBGT 10-18C	7 (39%)	12 (57%)	6 (35%)	7 (30%)	7 (41%)
Dry bulb temperature	11.6 (6.0)	12.4 (6.2)	18.9 (3.4)	11.7 (4.8)	13.2 (5.7)
Missing	0	1	1	1	1
Wet bulb temperature	7.6 (3.9)	8.6 (5.9)	14.9 (2.5)	7.6 (5.1)	9.9 (5.6)
Missing	0	1	1	1	1
Percent relative humidity	61 (21)	61 (11)	68 (16)	55 (18)	64 (16)
Missing	0	1	1	1	1
Black globe temperature	24 (9)	25 (6)	32 (8)	21 (6)	25 (7)
Missing	0	1	1	1	1
Solar radiation in Watts	654 (191)	460 (96)	679 (195)	401 (134)	437 (143)
Missing	0	1	1	1	1
Dew Point	3 (5)	5 (7)	12 (3)	3 (7)	6 (8)
Missing	0	1	1	1	1
Wind	12.0 (4.6)	8.2 (3.3)	9.2 (2.9)	11.2 (4.7)	8.8 (3.3)
Missing	0	1	1	1	1
WBGT	11.3 (4.6)	12.1 (5.9)	18.6 (3.3)	10.7 (5.0)	13.3 (5.6)
Missing	0	1	1	1	1
Air Quality Index	42 (15)	40 (13)	37 (15)	33 (14)	35 (15)

[1] Mean (SD) for continuous; n (%) for categorical

In addition, `Table 3` presents a summary statistics of all participants' characteristics by race. The New York City Marathon has most observations compared to others. The gender distribution is consistent across races with about 52% to 54% male participants. Participants for the Grandma's Marathon are relatively younger with an average age of 44 and Boston runners' %CR is lowest on average with the smallest standard deviation compared to other races' runners.

Table 3: Summary Table of Runner Characteristics

Characteristic	Race				
	Boston Marathon, N = 2,088	**Chicago Marathon, N = 2,553**	**Grandma's Marathon, N = 2,000**	**New York City Marathon, N = 2,930**	**Twin Cities Marathon, N = 1,993**
Gender					
Female	984 (47%)	1,210 (47%)	934 (47%)	1,402 (48%)	922 (46%)
Male	1,104 (53%)	1,343 (53%)	1,066 (53%)	1,528 (52%)	1,071 (54%)
Age	47 (17)	46 (18)	44 (18)	50 (19)	45 (17)

Table 3: Summary Table of Runner Characteristics *(continued)*

Characteristic	Race				
	Boston Marathon, N = 2,088	**Chicago Marathon, N = 2,553**	**Grandma's Marathon, N = 2,000**	**New York City Marathon, N = 2,930**	**Twin Cities Marathon, N = 1,993**
Percent Off Course Record (%CR)	41 (34)	52 (47)	48 (40)	55 (56)	46 (37)

[1] Mean (SD) for continuous; n (%) for categorical

Aim 1: Effect of Incresing Age in Men and Women

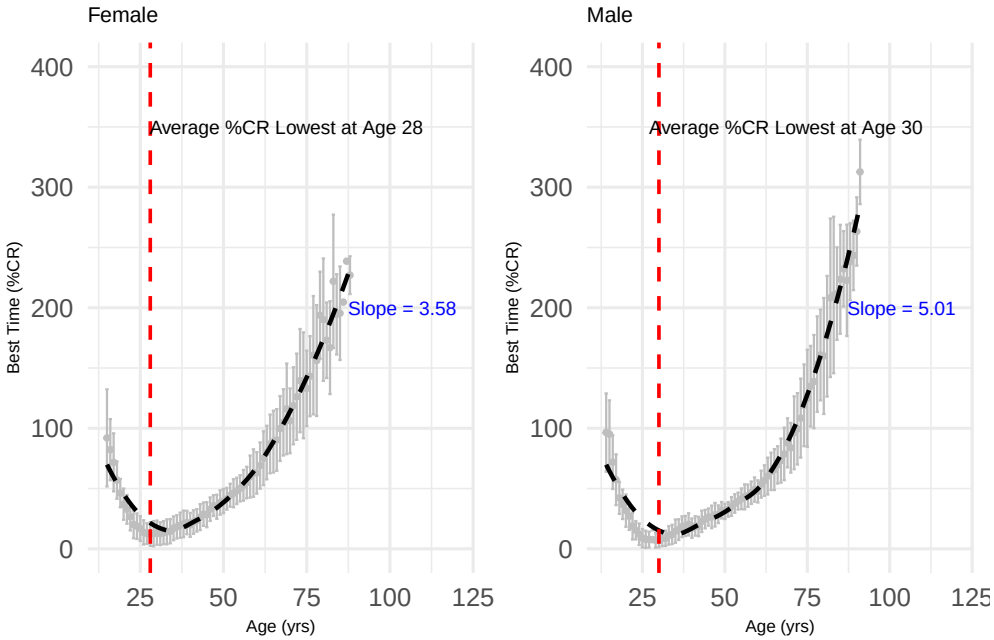
We first investigate the relationship between age and performance by gender. `Table 4` presents the summary of participants' performance based on their age group and gender. The younger and highest age groups have fewer observations due to the limited number of participants at these extreme ages. The lower-mid age group participants had the best performance, with both genders exhibiting the lowest %CR values across all aging groups. People with younger, Mid, and upper-mid age have relatively worse performance compared to lower-mid age people. Older people (Highest age) show the largest deviation from the course record, with an average %CR at around 130-140, and the %CR value for this group has the widest range. In addition, female runners generally have worse performance (higher %CR value) compared to males runners within the same age range.

Table 4: Summary of Marathon Performance by Age Group and Sex

Age Group	Sex	N	Min Performance	Mean Performance	Median Performance	Max Performance
Younger	Female	788	-0.385	40.840	38.051	211.095
Younger	Male	834	-1.074	36.119	31.196	159.535
Lower-Mid Age	Female	1440	-1.816	15.547	14.218	50.301
Lower-Mid Age	Male	1440	-2.251	11.645	9.322	50.179
Mid Age	Female	1440	-1.419	34.064	33.975	76.215
Mid Age	Male	1440	1.243	27.992	28.445	89.271
Upper-Mid Age	Female	1346	21.154	75.345	69.512	273.824
Upper-Mid Age	Male	1435	11.663	58.746	55.054	153.190
Highest Age	Female	438	36.187	138.609	129.911	336.347
Highest Age	Male	963	48.375	132.480	120.909	419.958

Additionally, in `Figure 2`, both genders' plots exhibit a U-shaped curve which further explains the relationship between age and performance. The performance continuously improves as age increases (%CR decreases as age increases), achieve the optimal performance (lowest %CR), and then gradually declines as age increases further (%CR increases as age increases). Both genders reach their optimal performance during their Lower-Mid age where female participants' average %CR achieves the lowest value at age of 28 and male participants achieves the lowest average %CR at age 30. After reaching their optimal performance, both genders exhibits sharp decline as age increases. Male participants show more extreme decline (with a slope of 5.01) compared to female participants (with a slope of 3.58) especially during their upper-mid and highest age, end with a maximum average percent off course record (CR%) exceeding 300.

Figure 2: Overall Performance vs. Age by Gender



From these results, we found that age does affect the marathon performance for both genders. Runners' performance gradually enhances from their younger age to lower-mid age, reaching their peak at age around 30, after which performance gradually declines. There are also gender differences that females usually perform worse compared to males within the same age group. In addition, male runners with higher age have more pronounced decline in performance compared to female runners.

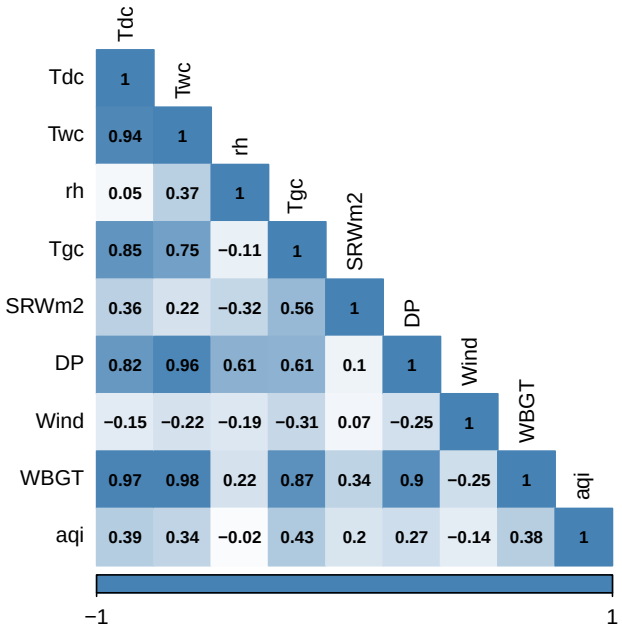
Aim 2: Impact of Environmental Conditions across Age and Gender

Environmental factors such as humidity, WBGT, and wind speed are also suspected to be key influences on marathon performance. To investigate the effects of these environmental condition characteristics on marathon performance, we first examine the correlations between key environmental variables and performance metrics by the correlation plot in `Figure 3`. Strongly correlated variables, either positively or negatively, are in blue, and weakly correlated variables are presented in white. Observing `Figure 3`, WBGT, dry bulb, wet bulb, and black global temperature show strong positive correlation with each other. This is reasonable since WBGT is calculated with those temperature variables using the following formula:

$$WBGT = (0.7 * Twc) + (0.2 * Tgc) + (0.1 * Tdc)$$

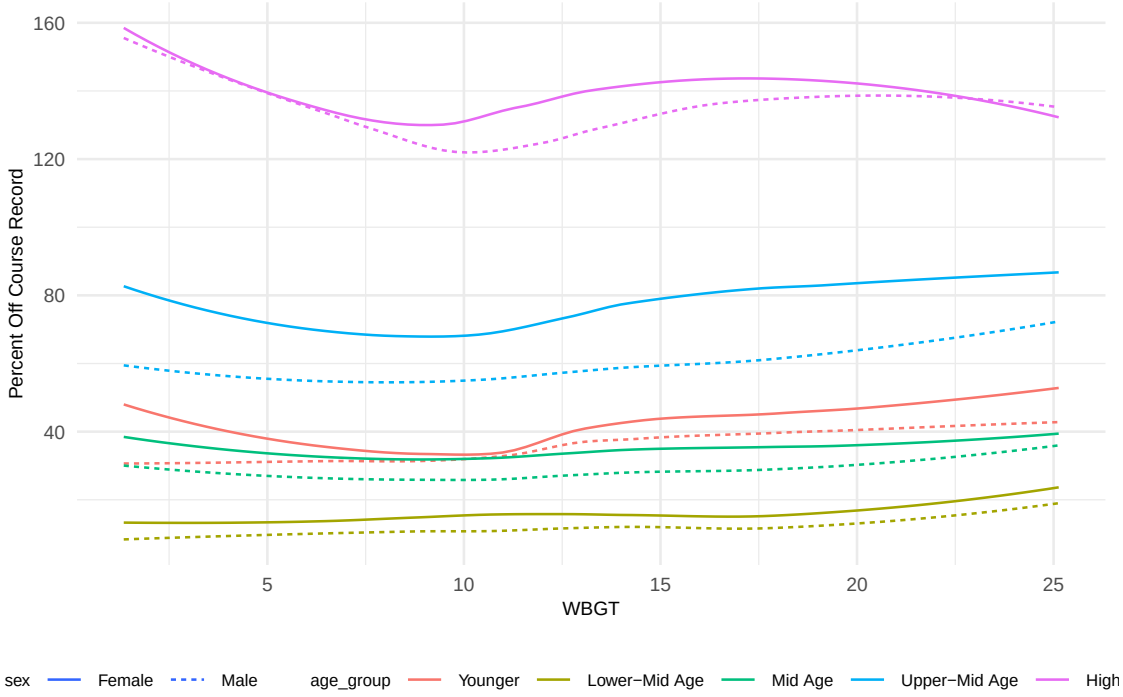
All other environmental characteristics have moderate or week relationship among each other. For example, the relative humidity (`rh`) is weakly and negatively correlated with `Tdc` and `Tgc` with correlation value of -0.01.

Figure 3: Correlation Plot among Environmental Condition Characteristics



As mentioned previously, `WBGT` is calculated using all temperature variables and it is highly correlated with those variables. The strong correlation between `WBGT` and these variables suggests that `WBGT` alone can be a sufficient indicator of how temperature characteristics influence marathon outcomes. Higher `WBGT` values indicate higher risks of heat stress illness. `Figure 4` presents an illustration of how `WBGT` effects marathon performance across different age groups and genders, using `geom_smooth` to show the performance trend.

Figure 4: Runners' Performance vs. WBGT

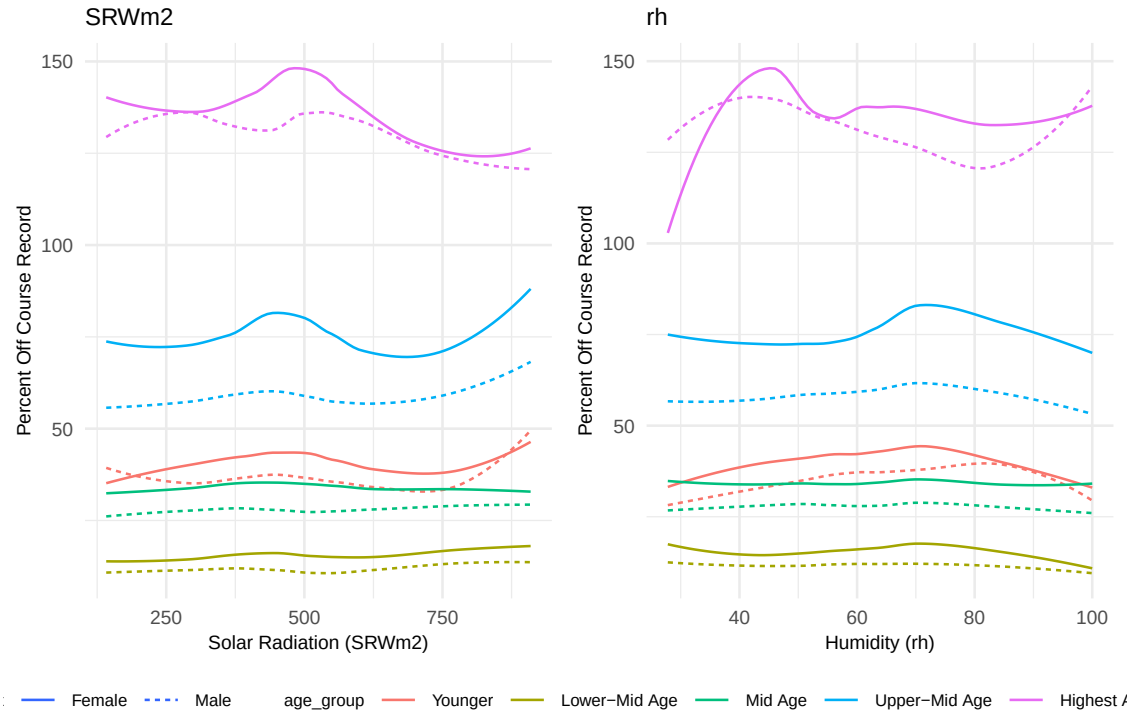


Runners in higher, upper-mid, and Younger age groups are more susceptible to the effects of increasing

WBGT that their performance fluctuates more as WBGT increases. Additionally, as WBGT increases, runners' performance generally worsens across all groups (%CR increase). Runners gradually perform better as WBGT value increase from $0°C$ to $10°C$, but their performance then declines when WBGT exceeds $10°C$. For older runners in the highest age group, this decline is more significant, leading to largest deviation from the course record. Additionally, male runners consistently perform better compared to female runners within the same age group as we mentioned earlier. Male runners in the younger, upper-mid, and highest age group show steeper decline in performance when WBGT is larger than $10°C$ compared to female younger runners. Although male runners perform better generally, they are more sensitive to higher WBGT levels in specific age groups.

Figure 5 and Figure 6 presents the relationship between other environmental condition characteristics and runners' performance across different genders and age groups.

Figure 5. Runners' Performance vs. Other Environmental Condition Factors (1)



Starting from Solar radiation (SRWm2), runners in the lower-mid and mid age groups seem to maintain a more stable performance as SRWm2 increase while younger, upper-mid, and highest age runners are more sensitive to the level of solar radiation, with their performance fluctuating more as solar radiation level increases. Moreover, female runners across groups show more stable performance compared to male runners that female runners are less likely to be influenced by solar radiation level. In general, solar radiation shows a slight increasing trend on people's performance with some fluctuations in the middle.

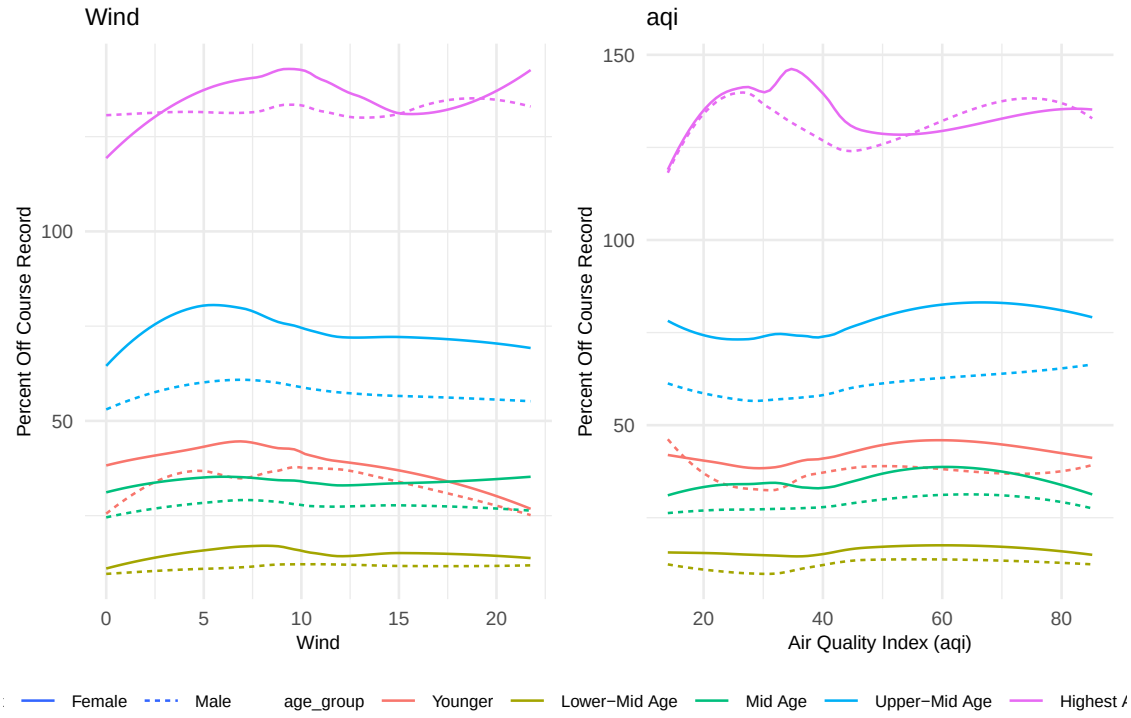
People in the younger, upper-mid, and highest age groups still show more fluctuated trend in relative Humidity (rh) as well. For most age groups, people show stable performance as relative humidity increase from 0% to around 60%. This indicates that moderate humid conditions would not negatively influence runner's performance. However, as humidity rises excess this range, runners start to have worse performance. Moreover, male runners across different age groups are much more sensitive to the humidity change compared to female runners.

Wind speed (Wind) does not exhibit significant trend on runners' performance across various age groups and genders. For most age groups, %CR shows only minor changes as wind speed increases, suggesting that wind speeds do not have significant impact on runners' performance that runners are able to adjust under various wind conditions and maintain their great performance. Male runners in the highest age group show relatively greater fluctuation since older runners would be more sensitive to environmental condition change. Moreover,

again, female runners present more stability in performance as wind speed varies as well.

Finally, runners show consistent performance as air quality index increases, except for the highest age group which experienced relatively more pronounced variation for both genders. In addition, male runners are more sensitive to air quality change compared to female runners within the same age group.

Figure 6. Runners' Performance vs. Other Environmental Condition Factors (2)



Thus, environmental condition characteristics like `WBGT`, solar radiation, and relative humidity do affect runner's performance. Older runners (upper-mid and highest age group) and younger runner (younger age group) are generally more sensitive to the environmental change. Females usually maintain higher stability through those environmental variations. Although wind speed and air quality do not present a pronounced effect on runners' performance, older runners still exhibits more sensitivity to these factors, while female runners continue to maintain more consistent performance compared to male runners.

Aim 3: Identify Important Weather Parameters

To investigate the significance of environmental condition characteristics on marathon performance, we first generated a correlation plot with all environmental condition characteristics and our response variable `CR_pct` shown in `Figure 7`. Observing the correlation plot, none of the factors exhibits strong correlation with runner's performance.

For further exploration, we fit a linear model and perform a backward model selection to present quantitative measures of significance. As shown in `Figure 3`, WBGT, dry bulb temperature, wet bulb temperature, and black globe temperature exhibit strong positive correlations, which is expected since WBGT is an index derived from these temperature variables. To address multicollinearity concerns, we include only WBGT as our temperature variable in the full model. Additionally, the model incorporates humidity, wind speed, solar radiation, air quality index, and interaction terms between age and these weather characteristics. We also include the quadratic term of age (age^2) in the full model to capture the nonlinear relationships between age and marathon performance. Then, use the full model, we perform a backward model selection procedure and the coefficient estimate results are shown in `Table 5`.

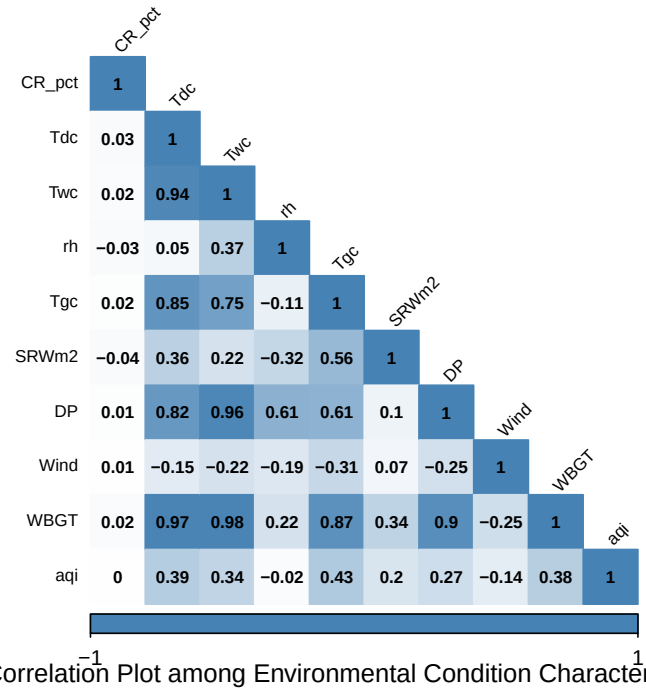


Figure 7: Correlation Plot among Environmental Condition Characteristics with %CR

For further exploration, we fit a linear model and perform a backward model selection to present quantitative measures of significance. As shown in `Figure 3`, WBGT, dry bulb temperature, wet bulb temperature, and black globe temperature exhibit strong positive correlations, which is expected since WBGT is an index derived from these temperature variables. To address multicollinearity concerns, we include only WBGT as our temperature variable in the full model. Additionally, the model incorporates humidity, wind speed, solar radiation, air quality index, and interaction terms between age and these weather characteristics. We also include the quadratic term of age (age^2) in the full model to capture the nonlinear relationships between age and marathon performance. Then, use the full model, we perform a backward model selection procedure and the coefficient estimate results are shown in `Table 5`.

Table 5: Coefficient Estimation of Best Model

Variable	Estimate	Standard Error	T Statistics	P Value
(Intercept)	114.2737167	3.6309472	31.472150	0.0000000
sexMale	-9.5066837	0.3770221	-25.215187	0.0000000
WBGT	0.2127369	0.1102679	1.929273	0.0537225
rh	0.0622261	0.0358442	1.736015	0.0825891
Wind	0.1548375	0.1355968	1.141897	0.2535217
SRWm2	0.0044593	0.0033216	1.342501	0.1794613
age	-5.5900725	0.0917711	-60.913185	0.0000000
I(age^2)	0.0798899	0.0006069	131.636557	0.0000000
WBGT:age	0.0078914	0.0022456	3.514112	0.0004430
rh:age	-0.0026038	0.0007202	-3.615237	0.0003014
Wind:age	-0.0045421	0.0026929	-1.686679	0.0916933
SRWm2:age	-0.0001933	0.0000677	-2.855631	0.0043031

Sex, age and the quadratic term of age exhibits significant estimates on the %CR. Age and age squared suggest a nonlinear relationship where performance generally worsens with age. Males, on average, perform better than females, with a %CR that is 9.51 percentage points lower compared to females. Among weather factors, WBGT shows a marginally significant positive association with %CR, meaning higher WBGT values (indicative of hotter, riskier conditions) are associated with worse performance. Moreover, the significant positive interaction between age and WBGT shows that the impact of WBGT on performance intensifies with increasing age, suggesting older runners are more adversely affected by high WBGT conditions compared

to younger runners. Relative humidity also demonstrates marginal significance ($p = 0.08$), suggesting that higher humidity worsens performance. The interaction between age and relative humidity has a significant negative estimate, indicating that the effect of humidity varies across age. Although wind speed and solar radiation do not have significant main effects, their significant interactions with age further underscore age-dependent influences of these environmental factors. Air Quality Index shows no significance as a main effect or interaction term. Notably, none of the interaction terms with sex are significant, emphasizing age as a critical factor in moderating the influence of weather on marathon outcomes.

Discussion

This report combines information from four datasets collected through the five major U.S. marathons over 15-20 years to investigate how age, sex, and environmental condition characteristics affect runner's performance and how effects of various weather conditions varies by age and gender. Through exploratory data analysis and statistical modeling, we concludes that age plays a significant role in runner's performance. Highest, Upper-mid, and younger, aged runners perform worse, especially for senior in the highest age group. Moreover, male runners show more sensitivity to age change compared to female runners, with a steeper decline in performance as age increases.

Our linear model reinforces our earlier findings about age, showing that both age and its quadratic term have significant effects, indicating a nonlinear relationship where performance generally declines with age. Males, with a negative coefficient estimate, perform better than females. Among weather parameters, WBGT has the most substantial impact on marathon performance, while other weather variables show either marginal significance or no significance. Notably, the influence of WBGT, humidity, wind speed, and solar radiation on performance varies across age, as evidenced by significant interactions between age and these weather variables.

One limitation of this report is the limited number of observations in the highest and youngest age groups which might lead to less reliability of our results. Further exploration with a larger sample size with more balanced aging group participants would help to present a more stable and reliable conclusion. In addition, participants of Boston Marathon may represent a more experienced sample, potentially limiting the generalizability of less experienced runners in our study. Moreover, as mentioned earlier, environmental factors are inherently associated with each other in reality, indicating the potential non-linear relationships or interaction effects between these variables. Further investigation of interaction and model selection might have a better and more comprehensive understanding of those factors' effects.

Reference

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.

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.

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

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.

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

Appendix

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)

# load necessary packages
library(tidyverse)
library(ggplot2)
library(visdat)
library(gtsummary)
library(kableExtra)
library(ggpubr)
library(gt)
library(car)
library(lme4)
library(lmerTest)
library(corrplot)
library(broom.mixed)
# set working directory
setwd("C:/Users/yingx/OneDrive/Desktop/Fall 2024/PHP 2550/Data/Project 1 Data/")

# read in data
data <- read.csv("project1.csv")
course_record <- read.csv("course_record.csv")
aqi_values <- read.csv("aqi_values.csv")
marathon_dates <-  read.csv("marathon_dates.csv")
# manage column names for easier understanding
colnames(data) <- c("race", "year", "sex", "flag", "age", "CR_pct", "Tdc",
                    "Twc", "rh", "Tgc", "SRWm2", "DP", "Wind", "WBGT")

# replace blank values in column `flag` to NA and factor it
data$flag <- case_when(data$flag == "" ~ NA, TRUE ~ data$flag)
data$flag <- as.factor(data$flag)

# Join course record data with the main data set
# re-code race and gender columns to fit with the main data's coding
course_record$Race <- case_when(course_record$Race == "B" ~ 0,
                                course_record$Race == "C" ~ 1,
                                course_record$Race == "NY" ~ 2,
                                course_record$Race == "TC" ~ 3,
                                course_record$Race == "D" ~ 4,
                                TRUE ~ NA)
course_record$Gender <- case_when(course_record$Gender == "M" ~ 1,
                                  course_record$Gender == "F" ~ 0,
                                  TRUE ~ NA)

# use left_join to merge the two data by gender, race, and year
data <- data %>%
  left_join(course_record, join_by("sex" == "Gender", "race" == "Race", "year" == "Year"))

# Join marathon_dates data with the main data set
# re-code race column in the main data to fit with the marathon_dates data
data$sex <- as.factor(data$sex)
data$race <- case_when(data$race == 0 ~ "Boston",
```

```
                        data$race == 1 ~ "Chicago",
                        data$race == 2 ~ "NYC",
                        data$race == 3 ~ "Twin Cities",
                        data$race == 4 ~ "Grandmas")

# use left_join to merge main data and marathon_dates by race and year
data <- data %>% left_join(marathon_dates, by = c("race" = "marathon", "year" = "year"))

# Join aqi_values data with the main data set
# calculate average air quality index by marathon and year
aqi_values_ave <- aqi_values %>%
  filter(sample_duration != "1 HOUR") %>%
  group_by(marathon, date_local) %>%
  summarize(ave_aqi = mean(aqi, na.rm = TRUE), .groups = "drop")

# manage date columns in main and aqi_values data for joining
data <- data %>% mutate(date = as.Date(date))
aqi_values_ave <- aqi_values_ave %>% mutate(date_local = as.Date(date_local))

# use left_join to merge the two data by race and date
data <- data %>%
  left_join(aqi_values_ave, by = c("race" = "marathon", "date" = "date_local"))

# manage column name of the new aqi column
colnames(data)[17] <- "aqi"

# re-code sex column
data$sex <- ifelse(data$sex == 1, "Male", "Female")

# factor and relevel flag column
data$flag <- factor(data$flag, levels = c("White", "Green", "Yellow", "Red", "Black", NA))
# # missing data heat map
# vis_dat(data) +
#   scale_fill_manual(values = rep("grey", 5)) +
#   theme(legend.position = "none",
#         plot.title = element_text(hjust = 0.5, size = 12)) +
#   ggtitle("Figure 1: Heat Map of Main Dataset")
# mutate a new column `age_group`
data$age_group <- cut(data$age, breaks = c(-Inf, 24, 39, 54, 69, Inf),
                      labels = c("Younger", "Lower-Mid Age", "Mid Age", "Upper-Mid Age", "Highest Age"))

# create a kable format table to present age group classification
age_grop_df <- data.frame(Group = c("Younger", "Lower-Mid Age", "Mid Age", "Upper-Mid Age", "Highest Age
                          age_range = c("Below 25", "25-39", "40-54", "55-69", "70+"))

knitr::kable(age_grop_df, col.names = c("Group", "Age Range"),
             caption = "Age Group Classifications for Marathon Participants") %>%
  kable_styling(latex_options = "HOLD_position",
                font_size = 10)
# merge two plots together
par(mfrow = c(1, 2), oma = c(0, 0, 3, 0))

# First histogram: original data
```

```r
hist(data$rh, col = "lightblue1", breaks = 13,
     xlab = "Relative Humidity",
     main = NULL, cex.main = 1.1, font.main = 1, cex.axis = 0.8, cex.lab = 0.9)

# modify the relative humidity values (those <= 1 are multiplied by 100)
data <- data %>%
  mutate(rh = ifelse(rh <= 1, rh * 100, rh))

# Second histogram: after modification
hist(data$rh, col = "lightblue1", breaks = 13,
     xlab = "Relative Humidity",
     main = NULL, cex.main = 1.1, font.main = 1, cex.axis = 0.8, cex.lab = 0.9)

mtext("Figure 1: Distribution of Relative Humidity Before and After Correction",
      outer = TRUE, cex = 1, adj = 0.5)
# create a new data.frame with one row for each race each year
single_year_data <- data %>%
  dplyr::select(-c("age", "sex", "CR_pct", "age_group", "date")) %>%
  mutate(race = case_when(race == "Boston" ~ "Boston Marathon",
                          race == "Chicago" ~ "Chicago Marathon",
                          race == "NYC" ~ "New York City Marathon",
                          race == "Twin Cities" ~ "Twin Cities Marathon",
                          race == "Grandmas" ~ "Grandma's Marathon",
                          TRUE ~ "Missing"),
         flag = case_when(flag == 'White' ~ "WBGT < 10C",
                          flag == 'Green' ~ "WBGT 10-18C",
                          flag == 'Yellow' ~ "WBGT > 18-23C",
                          flag == 'Red' ~ "WBGT > 23-28C",
                          TRUE ~ "Missing")) %>%
  distinct(race, year, .keep_all = TRUE)

# create a summary statistics of all environmental condition characteristics by race
single_year_summary <- single_year_data %>%
  dplyr::select(-c("year", "CR")) %>%
  tbl_summary(by = race,
              label = list(Tdc ~ "Dry bulb temperature",
                           Twc ~ "Wet bulb temperature",
                           rh ~ "Percent relative humidity",
                           Tgc ~ "Black globe temperature",
                           SRWm2 ~ "Solar radiation in Watts",
                           DP ~ "Dew Point",
                           aqi ~ "Air Quality Index"),
              statistic = all_continuous() ~ "{mean} ({sd})",
              missing = "ifany",
              missing_text = "Missing") %>%
  modify_spanning_header(update =  all_stat_cols() ~  "**Race**") %>%
  modify_footnote(update = all_stat_cols() ~ "Mean (SD) for continuous; n (%) for categorical") %>%
  bold_labels()

# convert the summary table to kable and print the result
single_year_summary %>%
  as_kable_extra(booktabs = TRUE, caption = "Summary Table of Weather Parameters",
                 longtable = TRUE, linesep = "") %>%
```

```
  kableExtra::kable_styling(font_size = 10,
                            latex_options = c("repeat_header", "HOLD_position", "scale_down"))%>%
  column_spec(1, width = "3cm") %>%
  column_spec(2, width = "2cm") %>%
  column_spec(3, width = "2cm") %>%
  column_spec(4, width = "2cm") %>%
  column_spec(5, width = "2cm") %>%
  column_spec(6, width = "2cm") %>%
  row_spec(0, bold = TRUE)
# create a summary statistics of all participants' characteristics by race
runners_tbl_summary <- data %>%
  dplyr::select(race, sex, age, CR_pct) %>%
  mutate(race = case_when(race == "Boston" ~ "Boston Marathon",
                          race == "Chicago" ~ "Chicago Marathon",
                          race == "NYC" ~ "New York City Marathon",
                          race == "Twin Cities" ~ "Twin Cities Marathon",
                          race == "Grandmas" ~ "Grandma's Marathon",
                          TRUE ~ "Missing")) %>%
  tbl_summary(by = race,
              label = list(sex ~ "Gender",
                           age ~ "Age",
                           CR_pct ~ "Percent Off Course Record (%CR)"),
              statistic = all_continuous() ~ "{mean} ({sd})",
              missing = "ifany",
              missing_text = "Missing") %>%
  modify_spanning_header(update =  all_stat_cols() ~  "**Race**") %>%
  modify_footnote(update = all_stat_cols() ~ "Mean (SD) for continuous; n (%) for categorical") %>%
  bold_labels()

# convert the summary table to kable and print the result
runners_tbl_summary %>%
  as_kable_extra(booktabs = TRUE, caption = "Summary Table of Runner Characteristics",
                 longtable = TRUE, linesep = "") %>%
  kableExtra::kable_styling(font_size = 10,
                            latex_options = c("repeat_header", "HOLD_position", "scale_down")) %>%
  column_spec(1, width = "3cm") %>%
  column_spec(2, width = "2cm") %>%
  column_spec(3, width = "2cm") %>%
  column_spec(4, width = "2cm") %>%
  column_spec(5, width = "2cm") %>%
  column_spec(6, width = "2cm") %>%
  row_spec(0, bold = TRUE)
# create a summary table by age group and sex
summary_table <- data %>%
  group_by(age_group, sex) %>%
  summarize(N = n(),
            min_performance = round(min(CR_pct, na.rm = TRUE), 3),
            mean_performance = round(mean(CR_pct, na.rm = TRUE), 3),
            median_performance = round(median(CR_pct, na.rm = TRUE), 3),
            max_performance = round(max(CR_pct, na.rm = TRUE), 3))

knitr::kable(summary_table,
             col.names = c("Age Group", "Sex", "N", "Min Performance",
```

```
                          "Mean Performance", "Median Performance", "Max Performance"),
             caption = "Summary of Marathon Performance by Age Group and Sex") %>%
  kable_styling(latex_options = "HOLD_position",
                font_size = 8)
# figures showing relationship between age and performance by gender
# summary female performance grouped by age for the plot
female_summary <- data %>%
  filter(sex == "Female") %>%
  group_by(age) %>%
  summarise(mean_CR = mean(CR_pct, na.rm = TRUE),
            se_CR = sd(CR_pct, na.rm = TRUE))

female_slope <- (female_summary$mean_CR[female_summary$age == 88] -
                 female_summary$mean_CR[female_summary$age == 28]) /
               (88 - 28)

# using geom_smooth with error bar setting age as x-axis and performance as response (female)
ageplot_female <- ggplot(female_summary, aes(x = age, y = mean_CR)) +
  geom_point(color = "grey", size = 1) +
  geom_errorbar(aes(ymin = mean_CR - se_CR, ymax = mean_CR + se_CR), width = 1, color = "grey") +
  geom_smooth(se = FALSE, color = "black", size = 1, method = "loess", linetype = 2) +
  geom_vline(xintercept = 28, linetype = "dashed", color = "red", size = 0.8) +
  annotate("text", x = 110, y = 350, label = paste("Average %CR Lowest at Age", 28),
           color = "black", hjust = 1, size = 3.2) +
  annotate("text", x = 120, y = 200, label = paste("Slope =", round(female_slope, 2)),
           color = "blue", hjust = 1, size = 3.2) +
  labs(title = "Female", x = "Age (yrs)", y = "Best Time (%CR)") +
  ylim(0, 400) +
  theme_minimal(base_size = 15) +
  theme(axis.title = element_text(size = 8),
        plot.title = element_text(size = 10))

# summary male performance grouped by age for the plot
male_summary <- data %>%
  filter(sex == "Male") %>%
  group_by(age) %>%
  summarise(mean_CR = mean(CR_pct, na.rm = TRUE),
            se_CR = sd(CR_pct, na.rm = TRUE))

male_slope <- (male_summary$mean_CR[male_summary$age == 91] -
                 male_summary$mean_CR[male_summary$age == 30]) /
               (91 - 30)

# using geom_smooth with error bar setting age as x-axis and performance as response (male)
ageplot_male <- ggplot(male_summary, aes(x = age, y = mean_CR)) +
  geom_point(color = "grey", size = 1) +
  geom_errorbar(aes(ymin = mean_CR - se_CR, ymax = mean_CR + se_CR), width = 1, color = "grey") +
  geom_smooth(se = FALSE, color = "black", size = 1, method = "loess", linetype = 2) +
  geom_vline(xintercept = 30, linetype = "dashed", color = "red", size = 0.8) +
  annotate("text", x = 110, y = 350, label = paste("Average %CR Lowest at Age", 30),
           color = "black", hjust = 1, size = 3.2) +
  annotate("text", x = 120, y = 200, label = paste("Slope =", round(male_slope, 2)),
           color = "blue", hjust = 1, size = 3.2) +
```

```
  labs(title = "Male", x = "Age (yrs)", y = "Best Time (%CR)") +
  ylim(0, 400) +
  theme_minimal(base_size = 15) +
  theme(axis.title = element_text(size = 8),
        plot.title = element_text(size = 10))

# merge the two plots together
fig1 <- ggarrange(ageplot_female, ageplot_male)
annotate_figure(fig1, top = text_grob("Figure 2: Overall Performance vs. Age by Gender",
                                      face = "bold", size = 10))
# create a correlation matrix among environmental condition factors
cor_matrix <- cor(data[,-c(1, 2, 3, 4, 5, 6, 15, 16, 18)], use = "complete.obs")

# correlation plot of environmental condition factors
corrplot(cor_matrix, method = "color", type = "lower",
         tl.col = "black", tl.cex = 0.8, addCoef.col = "black",
         number.cex = 0.7, col = colorRampPalette(c("steelblue", "white", "steelblue"))(200))
title("Figure 3: Correlation Plot among Environmental Condition Characteristics",
      cex.main = 0.9, line = 3)
# using geom_smooth to plot runners' performance vs. WBGT by gender and age groups
fig_WBGT <- ggplot(data, aes(x = WBGT, y = CR_pct, color = age_group, linetype = sex)) +
  geom_smooth(method = "loess", se = FALSE, size = 0.5) +
  labs(title = "WBGT",
       x = "WBGT", y = "Percent Off Course Record") +
  theme_minimal() +
  theme(axis.text = element_text(size = 8),
        axis.title=element_text(size = 8),
        legend.text=element_text(size = 7),
        legend.title=element_text(size = 7),
        plot.title = element_text(hjust = 0.5, size = 12),
        legend.position = "bottom") +
  ggtitle("Figure 4: Runners' Performance vs. WBGT") +
  guides(color = guide_legend(nrow = 1), linetype = guide_legend(nrow = 1))

fig_WBGT
# using geom_smooth to plot runners' performance vs. SRWm2 by gender and age groups
fig_SRWm2 <- ggplot(data, aes(x = SRWm2, y = CR_pct, color = age_group, linetype = sex)) +
  geom_smooth(method = "loess", se = FALSE, size = 0.5) +
  labs(title = "SRWm2",
       x = "Solar Radiation (SRWm2)", y = "Percent Off Course Record") +
  theme_minimal() +
  theme(axis.text = element_text(size = 8),
        axis.title=element_text(size = 8),
        legend.text=element_text(size = 7),
        legend.title=element_text(size = 7),
        plot.title = element_text(size = 10),
        legend.position = "bottom") +
  guides(color = guide_legend(nrow = 1), linetype = guide_legend(nrow = 1))

# using geom_smooth to plot runners' performance vs. rh by gender and age groups
fig_rh <- ggplot(data, aes(x = rh, y = CR_pct, color = age_group, linetype = sex)) +
  geom_smooth(method = "loess", se = FALSE, size = 0.5) +
  labs(title = "rh",
```

```
       x = "Humidity (rh)", y = "Percent Off Course Record") +
  theme_minimal() +
  theme(axis.text = element_text(size = 8),
        axis.title=element_text(size = 8),
        legend.text=element_text(size = 7),
        legend.title=element_text(size = 7),
        plot.title = element_text(size = 10),
        legend.position = "bottom") +
  guides(color = guide_legend(nrow = 1), linetype = guide_legend(nrow = 1))

# using geom_smooth to plot runners' performance vs. Wind by gender and age groups
fig_wind <- ggplot(data, aes(x = Wind, y = CR_pct, color = age_group, linetype = sex)) +
  geom_smooth(method = "loess", se = FALSE, size = 0.5) +
  labs(title = "Wind",
       x = "Wind", y = "Percent Off Course Record") +
  theme_minimal() +
  theme(axis.text = element_text(size = 8),
        axis.title=element_text(size = 8),
        legend.text=element_text(size = 7),
        legend.title=element_text(size = 7),
        plot.title = element_text(size = 10),
        legend.position = "bottom") +
  guides(color = guide_legend(nrow = 1), linetype = guide_legend(nrow = 1))

# using geom_smooth to plot runners' performance vs. aqi by gender and age groups
fig_aqi <- ggplot(data, aes(x = aqi, y = CR_pct, color = age_group, linetype = sex)) +
  geom_smooth(method = "loess", se = FALSE, size = 0.5) +
  labs(title = "aqi",
       x = "Air Quality Index (aqi)", y = "Percent Off Course Record") +
  theme_minimal() +
  theme(axis.text = element_text(size = 8),
        axis.title=element_text(size = 8),
        legend.text=element_text(size = 7),
        legend.title=element_text(size = 7),
        plot.title = element_text(size = 10),
        legend.position = "bottom") +
  guides(color = guide_legend(nrow = 1), linetype = guide_legend(nrow = 1))
# merge the plots
plot_mix1 <- ggarrange(fig_SRWm2, fig_rh,
                       ncol = 2, common.legend = TRUE, legend = "bottom")

# annote the mixed plot with title
annotate_figure(plot_mix1,
                top = text_grob("Figure 5. Runners' Performance vs. Other Environmental Condition Factor
                                size = 12, hjust = 0.5))
# merge the plots
plot_mix2 <- ggarrange(fig_wind, fig_aqi,
                       ncol = 2, common.legend = TRUE, legend = "bottom")

# annote the mixed plot with title
annotate_figure(plot_mix2,
                top = text_grob("Figure 6. Runners' Performance vs. Other Environmental Condition Factor
                                size = 12, hjust = 0.5))
```

```r
# create a correlation matrix among environmental condition factors
cor_matrix_CR <- cor(data[,-c(1, 2, 3, 4, 5, 15, 16, 18)], use = "complete.obs")

# correlation plot of environmental condition factors
corrplot(cor_matrix_CR, method = "color", type = "lower",
         tl.col = "black", tl.cex = 0.7,  tl.srt = 45, addCoef.col = "black",
         number.cex = 0.7, col = colorRampPalette(c("steelblue", "white", "steelblue"))(200))
mtext("Figure 7: Correlation Plot among Environmental Condition Characteristics with %CR",
      side = 1, cex.main = 0.9, line = 4)
# fit a full model with interaction terms
linear_model <- lm(CR_pct ~ sex * (WBGT + rh + Wind + SRWm2 + aqi) + age * (WBGT + rh + Wind + SRWm2 + a

# perform backward selection
selection_model <- step(linear_model, direction = "backward", trace = 0)
coef_result <- summary(selection_model)$coefficients

# convert the summary table to kable and print the result
coef_result %>%
  kable(booktabs = TRUE, caption = "Coefficient Estimation of Best Model",
        col.names = c("Variable", "Estimate", "Standard Error", "T Statistics", "P Value")) %>%
  kable_styling(font_size = 7, latex_options = c("repeat_header", "HOLD_position", "scale_down"))
```