Exploring the Relationships between Environmental Conditions and Marathon Performance

Yunan Chen

2024-10-06

Abstract

Environmental conditions are critical for endurance sports such as marathons and may affect a runner's performance. The effects may vary by gender and age. Utilizing data from five marathons collected over a 17 to 23-year period, this report explores the relationship between environmental variables and marathon performance across gender and age. The findings suggest a "U-shape" relationship between age and performance. Weather conditions were more likely to affect performance in the elderly than in the youth and adults. Limitations such as data coding errors, missing data, lack of information on extreme weather conditions, and the non-linear relationship emphasize the need for further research and examination.

Introduction

The implications of weather conditions on athletics performance in endurance exercises such as marathons have raised considerable attention over the past few years. Among all weather-related factors, temperature (i.e., heat or cold), relative humidity (i.e., dry or humid), wind speed, solar radiation, and air quality were considered to have the largest potential effect on undermining athletic performance. Previous studies have found that the decline in performance is associated with warmer temperatures. Older people, in particular, suffer from impaired thermoregulation and hence weakened ability to dissipate heat, which may further exacerbate the effects of warmer temperatures. In addition, there are well-documented sex differences in endurance performance and physiological processes related to thermoregulation.

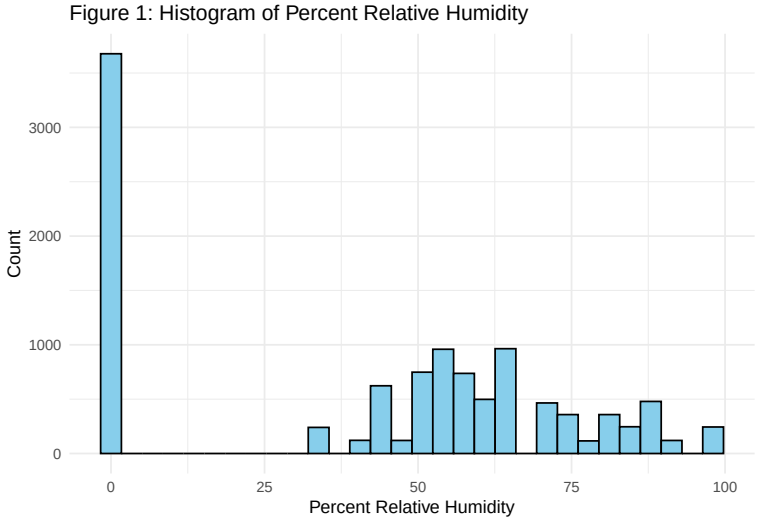
This report explores the relationship between age and marathon performance in men and women. Exploring the impact of environmental conditions including temperature, relative humidity, solar radiation, and wind speed on marathon performance, and whether the impact differs across age and gender. Identifying the weather conditions that have the largest impact on marathon performance.

Data Collection

Data used in this report was provided by Dr. Brett Romano Ely and Dr. Matthew Ely. Data collected in the study were coded into four seperate datasets: `project 1`, `aqi_values`, `course_record`, and `marathon_dates`. The major dataset is `project 1` which included the top single-age performances from five major marathons across 17-23 years from men and women, with detailed information on weather conditions for each marathon. The performance variable `%CR` was coded as the percent of the current course record. The `%CR` value indicates how many percentage points slower the participant's performance was than the couse record. A low `%CR` value indicates that the performance is close to the current course record. Since Wet Bulb Globe Temperature (WBGT) is calculated as WBGT $= 0.7 \cdot$ Wet Bulb Globe Temperature $+ 0.2 \cdot$ Black Bulb Globe Temperature $+ 0.1 \cdot$ White Bulb Globe Temperature, and the variable `Flag` was grouped into five categories based on the value of WBGT and risk of heat illness, the temperature-related variables in the data are expected to be highly correlated. The other three datasets `aqi_values`, `course_record` and `marathon_dates` contain information on air quality, course records, and dates of each marathon respectively.

Data Pre-processing

The raw dataset `project 1` contained 11,564 observations and 14 variables. To get a more comprehensive dataset, the average air quality information of each marathon was calculated using the variable `AQI` in dataset `aqi_values` and was merged to the major data set `project 1` by years and marathon. Some values of the humidity variable appeared to be reported incorrectly. By looking at the histogram in percent relative humidity, we can see that more than 3500 observations fall around 0, which is highly unlikely to be the case since the variable should be in percentage (`Figure 1`). Therefore, we modified the values that were less than or equal to 1 by multiplying them by 100. In addition, in the original dataset, missing values in the categorical variable `Flag` were coded as empty cells, which would cause R to interpret them as one of the categories by default. Therefore, these values were recoded as NA.



Based on the age distribution, a categorical variable `age_grp` with five levels was created for later exploratory analyses. The five levels were < 25 years, 25-39 years, 40-54 years, 55-69 years, and >= 70 years. The number of participants in each age group is shown in `Table 1`. There were more participants in the middle-aged age group compared to the oldest and youngest age groups.

Table 1: Number of Participants by Age Group and Gender

	Gender	
Characteristic	**Female**, N = 5,452	**Male**, N = 6,112
Age Group		
< 25 yrs	788	834
25-39 yrs	1,440	1,440
40-54 yrs	1,440	1,440
55-69 yrs	1,346	1,435
>= 70 yrs	438	963

`Table 2` shows all weather-related information and the course record by marathon. We can see that data were collected on five marathons: Boston, Chicago, New York City, Twin Cities, and Grandma's Marathon. The "N" next to the marathon's name stands for the number of years the data collection crossed. For example, data on the Boston Marathon were collected across 18 years. Information on weather conditions in Chicago, New York, the Twin Cities, and Grandma's Marathon was missing for one of the years. The 2011 Chicago, New York and Twin Cities Marathons and the 2012 Grandma's Marathon appeared to be missing all weather-related variables. Further research may be needed to determine why this happened.

Table 2: Summary Statistics of Weather Condistions by Race

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					
WBGT <10C	9 (50%)	6 (29%)	0 (0%)	11 (48%)	5 (29%)
WBGT 10-18C	7 (39%)	12 (57%)	6 (35%)	7 (30%)	7 (41%)
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%)
Missing	0 (0%)	1 (4.8%)	1 (5.9%)	1 (4.3%)	1 (5.9%)
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
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
Dew Point	3 (5)	5 (7)	12 (3)	3 (7)	6 (8)
Missing	0	1	1	1	1
Course Record in seconds	8,445 (47)	8,312 (111)	8,849 (60)	8,611 (64)	8,819 (23)
Average Air Quality	42 (15)	40 (13)	37 (15)	33 (14)	35 (15)

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

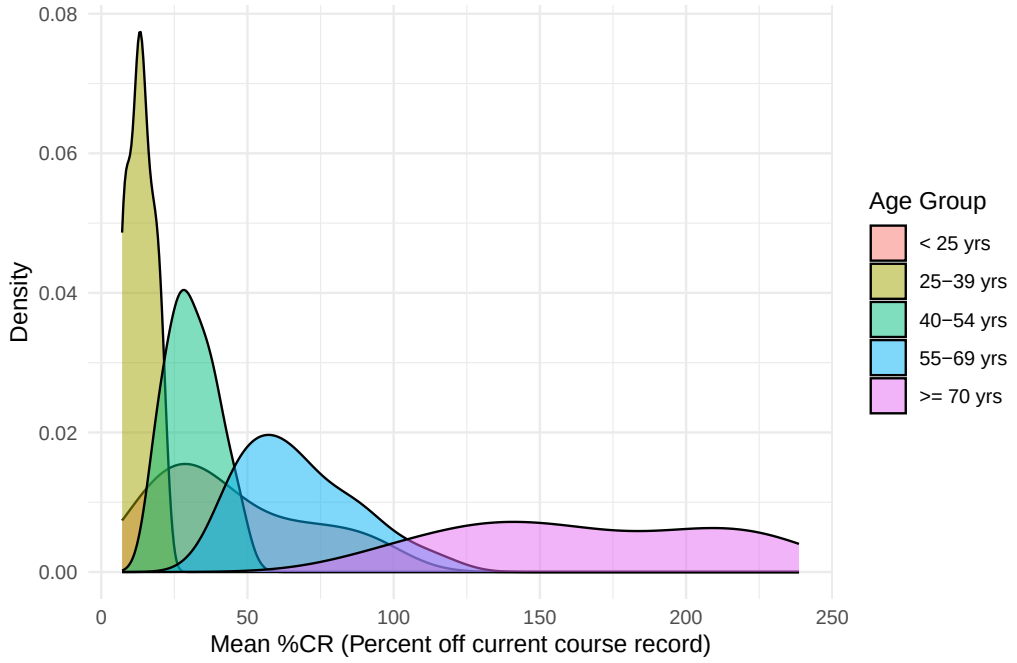
Marathon Performance by Age and Gender

We started examining the effects of increasing age on marathon performance in both men and women by getting a general idea of how the distribution of the average percentage off the current record (`%CR`) would vary across age groups. In this data set, the minimum age was 14 years, and the maximum age was 91 years.

`Figure 2` shows the distribution of `%CR` for the five age groups. There were differences in the shape of the distributions across age groups, with the 25-39 age group having the narrowest distribution and the >=70 age group having the most spread out distribution. This suggests there was less variation in the performance of the participants aged between 25 and 39 compared to the older and younger age groups. In addition, the mean value of `%CR` appeared to be the smallest in the 25-39 age group, which suggests that participants in this age group tended to perform better compared to the other age groups. We may expect that the top performance was achieved by the participants in this age group. The >=70 and 55-69 age groups had the

largest and second-largest variations in performance.

Figure 2: Distribution in mean %CR by Age Group



To further explore whether the patterns we saw in the distribution plot apply to all genders, `Table 3` was created to compare the mean `%CR` across age groups and genders. For both males and females, the lowest average `%CR` was found in the 25-39 age group, with a mean of 11.64 for men and 15.55 for women This suggests that men and women in the 25-39 age group had, on average, the best performance, with the times closest to the course record. This is consistent with what was seen in the distribution plot. In the oldest age group (>=70), the mean `%CR` was over 100%, indicating that older participants were more than twice as slow as the course record. Across all age groups, men tended to do slightly better than women, with `%CR` in men lower by 4 to 7 points compared to women. Trends in performance were similar for men and women in all age groups, with a decline in performance below and above the 25-39 age group.

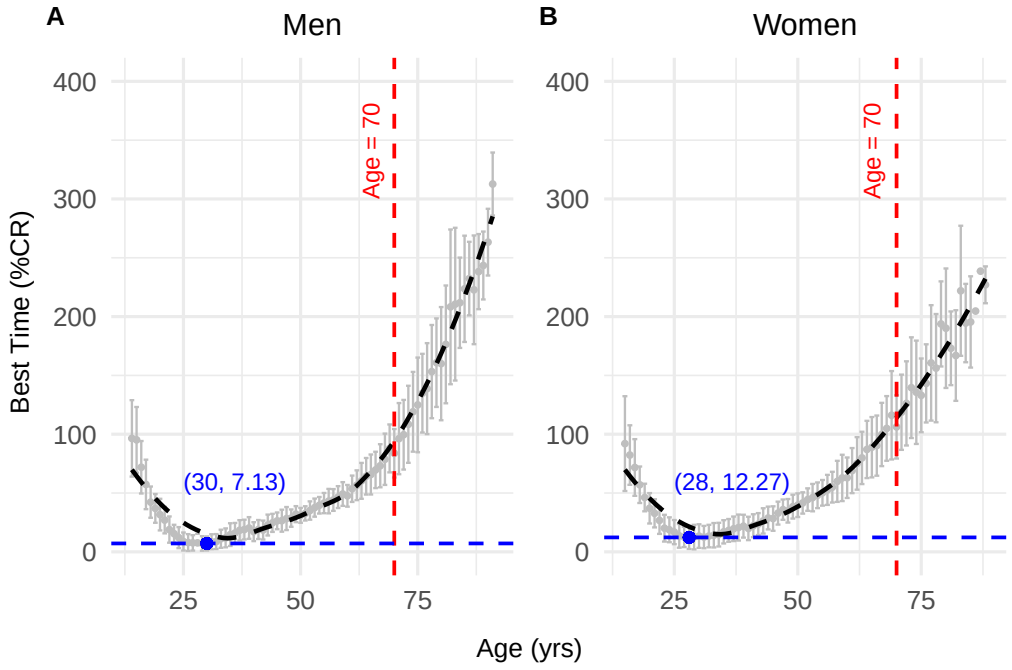
Table 3: Summary Statistics of %CR by Gender, Mean (Q1, Q3)

Characteristic	Men, N=6,112	Women, N=5,452
CR% (by Age Group)		
< 25 yrs	36.12 (16.9, 44.77)	40.84 (23.93, 52.95)
25-39 yrs	11.64 (4.42, 17.62)	15.55 (6.58, 22.97)
40-54 yrs	27.99 (21.69, 34.6)	34.06 (24.85, 43.07)
55-69 yrs	58.75 (44.36, 68.37)	75.35 (54.47, 89.25)
>= 70 yrs	132.48 (90.13, 164.09)	138.61 (103.96, 166.44)

`Figure 3` visualizes the relationship between age and `%CR`, with the mean and standard error of the `%CR` at each age indicated in gray and the overall trend indicated by the black dashed curve. Both men and women show a "U-shaped" curve as age increases. There was a decrease in `%CR` up to a certain age and an increase thereafter. The best results for males occurred around age 30 and for women around age 28, while poorer records occur at younger and older ages. The blue dots in the graph denote the best performances at the respective ages. For example, the best average performance for men occurred at age 30, with a performance 7.13% slower than the course record. The slope of the curve became more stepped after the age of 50, which

suggests that the decline was initially gradual but became more pronounced after the age of 50, with an even greater decline after the age of 70. In addition, the decline after age 70 appeared to be greater for men than for women. This may indicate that age has a greater impact on men's performance in their later years compared to women's.

Figure 3: Interaction between %CR and Age by Gender



Potential Enviornmental Variables

The purpose of Dr. Eli's study was to examine the effects of environmental conditions (including temperature, humidity, solar radiation, and wind) on lifetime marathon performance in both males and females. In the following analysis, we will focus on only four weather-related variables, namely WBGT, percent humidity, solar radiation, and wind.

Wet Bulb Globe Temperature (WBGT)

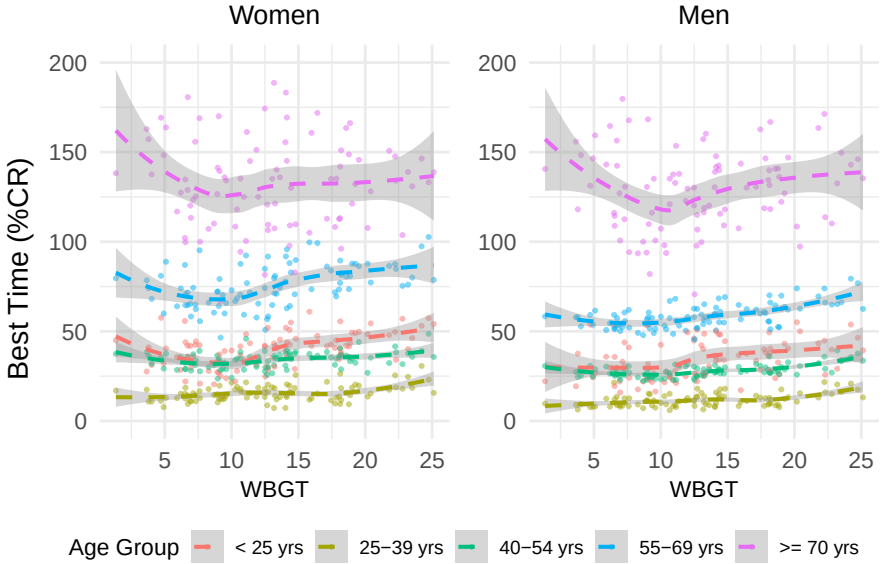
Recalling the description of the variables Wet Bulb Globe Temperature (WBGT), it is a measure of temperature taking into account dry, wet, and global temperatures and was used to explore the relationship between temperature and marathon performance. High WBGT values would lead to high levels of environmental heat stress, meaning that it will be more challenging for the body to regulate its internal temperature. And this can make physical activities, like running a marathon, much more taxing and potentially dangerous.

`Figure 4` visualizes the relationship between WBGT and the marathon performance (`%CR`) for different age groups of men and women. The dots in the graph indicate the average performance of different age groups at different WBGT levels. The plot was organized into two panels, one for women and the other for men. Smoothed lines were fitted to the data points to get an overall picture of the trends.

The fitted lines are always at the bottom for the 25-39 age group and always at the top for the >=70 age group, suggesting that regardless of changes in the WBGT conditions, on average, the 25-39 age group always outperforms the other age groups and the >=70 age group always performs the worst on average. This statement holds true for both men and women.

For both men and women in all age groups,`%CR` decreased with increasing WGBT when WGBT was less than 10, while `%CR` increased with increasing WGBT when WBGT was greater than 10. This suggests that when WGBT is less than 10, participants performed better as WGBT increases, while when WGBT is greater than 10, participants performed worse as WGBT increases. The rise and fall in `%CR` with increasing WBGT was clearest among participants in the oldest age group (>=70), showing the greatest change in performance and indicating that they were more likely to be affected by the WBGT. The fitted lines for the younger group are flatter, which means that their performance would not change much depending on the change in the WBGT.

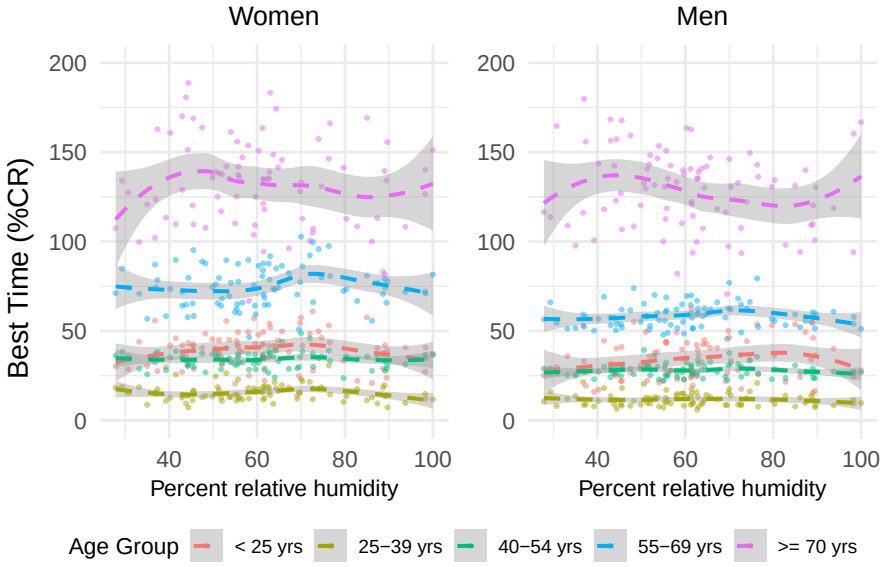
Figure 4: Interaction between %CR and WBGT with Age by Gender



Percent Humidity

`Figure 5` visualizes the relationship between average humidity and `%CR` for different age groups of men and women. The oldest age group (>=70 yrs) had the curviest fitted lines, indicating that this age group had the greatest fluctuation in marathon performance (`%CR`) with humidity. Marathon performance for this age group first increased from about 0 to 50, then became stable between 50 and 80, and finally increased after 80. The 55-69 age group showed a similar trend, but the fluctuations were not as pronounced as for the >=70 age group. There was a steady trend in the performance of participants in all age groups between 50 and 80 humidity, indicating that the performance did not change much when the relative humidity percentage fall into this range. The much flatter fitted line for the younger age groups suggests that participants in these age groups are less affected by humidity. The relationship between humidity and marathon performance was similar for males and females.

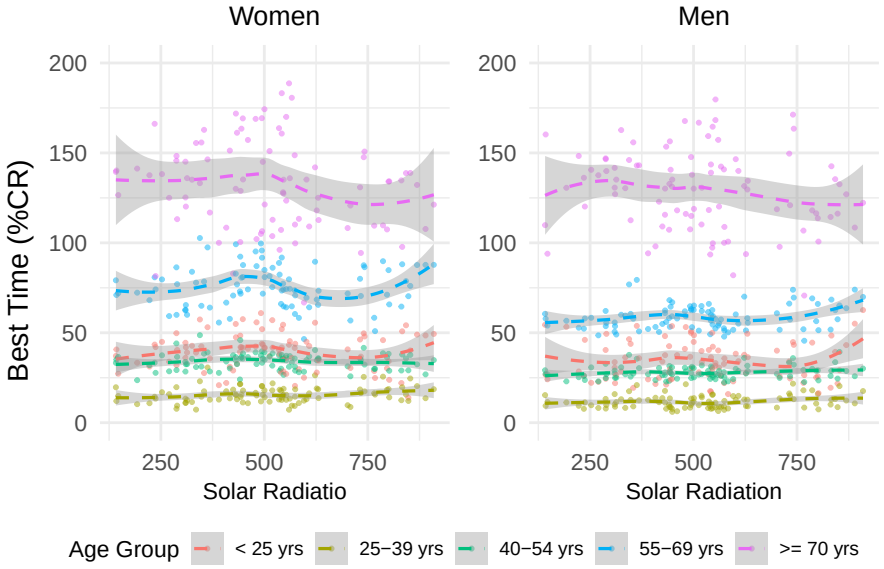
Figure 5: Interaction between %CR and Humidity with Age by Gender



Solar Radiation

`Figure 6` visualizes the relationship between solar radiation and marathon performance ("%CR") for different age groups of men and women. Firstly, according to the position of the fitted line, participants in the age group 25-39 years always performed the best regardless of the variation of solar radiation, followed by the 40-54, <25, 55-69 and >=70 years age groups. When solar radiation is below 500, the fitted lines are relatively flat for the <25, 25-39, and 40-54 age groups for both sexes. This suggests that when solar radiation is in the range of 0-500, it hardly affects the performance of runners aged 14-54 years. There is a "U-shaped" relationship between solar radiation and the performance of females aged 55-69 and >=70 years. This suggests that for female runners in these age groups, performance will first decline and then rise above the 500 solar radiation threshold.

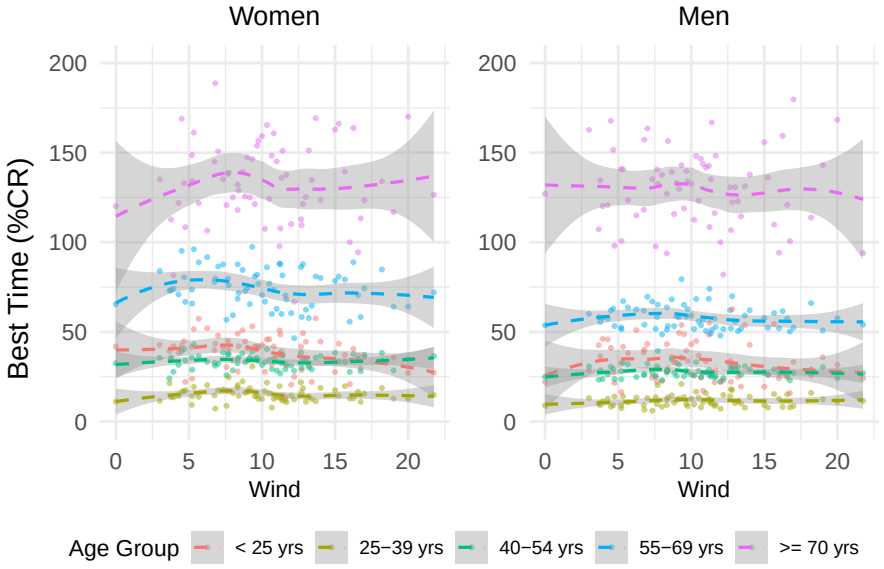
Figure 6: Interaction between %CR and Solar Radiation with Age by Gender



Wind

`Figure 7` visualizes the relationship between Wind and the marathon performance (`%CR`) for different age groups of men and women. On average, men and women in the 25-39 age group always outperformed the other age groups, while men and women in the >=70 age group always performed the worst. The fitted line is more curvy for the oldest age group of females as compared to males. This suggests that changes in wind speed introduce greater variation in marathon performance for older women compared to men. In the oldest age group of females, `%CR` increased and performance decreased when wind speed increased from 0 to 7.5. This was only seen in females, while males had no significant change in `%CR` in this wind speed range. There was no significant change in performance for either sex in the younger group as the wind changed.

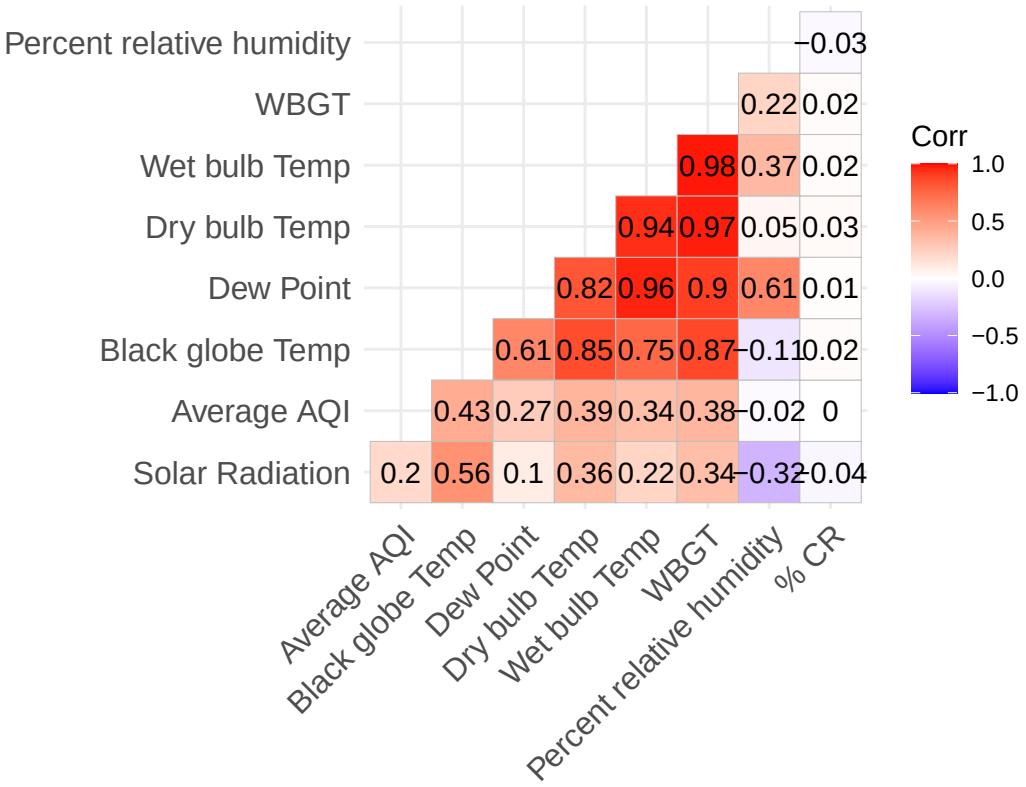
Figure 7: Interaction between %CR and Wind with Age by Gender



Importance Variables

`Figure 8` presents the correlation between weather-related variables and marathon performance (`%CR`). Positive correlations are shown in red and negative correlations are shown in blue. The higher the correlation, the darker the color. The last column of the correlation plot shows that the correlations between performance `%CR` and environmental variables are all around 0. This indicates that there was no linear relationship between performance and weather conditions. The Wet Bulb Globe Temperature (WBGT) variable is correlated with Dry Bulb Temperature, Wet Bulb Temperature, Black Bulb Temperature, and Dew Point, with a correlation close to 1. This suggests to include only one of these four variables when modeling performance to avoid co-linearity. As a result, the WBGT, humidity, solar radiation, wind, and average air quality will be included as covariates in the model to determine their effects on marathon performance.

Figure 8: Correlation Matrix of Environmental Variables



To quantify the association between weather-related variables and marathon performance, a random intercept linear mixed model was used. The model included the covariates WBGT, humidity, solar radiation, wind, and average air quality, with intercepts varying by age. The random intercept component was utilized because there was a relationship between age and performance as can be seen in `Figure 3`, and we wanted to capture the variation in performance by age in the model. The model has the formula

$$\%\text{CR}_{ij} = \beta_0 + \beta_1 \cdot \text{WBGT}_{ij} + \beta_2 \cdot \text{Humidity}_{ij} + \beta_3 \cdot \text{Solar Radiation}_{ij} + \beta_4 \cdot \text{Wind}_{ij} + \beta_5 \cdot \text{Ave AQI}_{ij} + u_i + \epsilon_{ij}$$

, where ij represents the jth observation in the ith age group (

$$i \in (14, 15, \cdots, 91)$$

).

`Tables 4` and `5` display the coefficient estimates, standard error, and p-values for the fixed and random effects in the previously described model. The variables WBGT, percent humidity, and solar radiation appeared

to be significant at the 0.05 level of significance, suggesting that all three factors are likely to have an effect on marathon performance regardless of age. However, the coefficient estimates for each of these variables are small, close to 0, and while the effects are significant, they do not result in a large change in marathon performance. For example, a one-point increase in WBGT would only result in a 0.5 increase in %CR of performance, adjusted for age. Variable wind speed and average AQI appeared to be insignificant. This model has limitations. Based on `Figure 3`, we can see that the relationship between marathon performance and age is nonlinear. The current model does not accommodate this structure. Model selection is expected to be carried out in further regression analyses.

Table 4: Fixed Effects of the Model

Term	Estimate	Standard Error	P-value
(Intercept)	79.0589876	8.4219228	0.0000000
WBGT	0.5326204	0.0398633	0.0000000
Percent Humidity	-0.0647072	0.0122432	0.0000001
Solar Radiation	-0.0048888	0.0011287	0.0000150
Wind	-0.0360238	0.0460462	0.4340321
Average AQI	0.0017624	0.0135754	0.8967092

Table 5: Random Effect of the Model

Group	Term	SD
Age	Intercept	73.55379
Residual		18.68690

Discussion

Based on previous research on marathon performance, we expected performance to vary by age, gender, environmental conditions, or combinations of them. The exploratory analyses in this report suggest a "U-shaped" relationship between marathon performance and age, whereby performance initially rises, peaks at a certain age, and then declines. On average, women reach their highest performance at age 28 and men at age 30. Environmental conditions considered in the analysis included WBGT, humidity, solar radiation, and wind speed. Regardless of these environmental conditions, the 25-39 age group always outperformed the other age groups, and the oldest age group always had the highest `%CR`, meaning that they took the longest to complete the race. There were also some patterns between these four environmental conditions and the performance (`%CR`) of the oldest age group. For example, when the WBGT was increased from 0 to 10, the performance of both males and females in the oldest age group increased. However, no clear pattern was observed from the interaction plots, suggesting that weather conditions tended to affect younger participants.

There are several limitations in this study. First, the number of observations for the youngest and oldest age groups is relatively small compared to the middle-aged groups (`Table 1`). This may lead to a reduction in statistical power for the youngest and the oldest age groups. The findings for the youngest and oldest groups may not be as generalizable to the broader population in these age ranges. Second, this data contains little information on marathon performance in extreme weather conditions. This reduces the ability to predict how environmental factors, such as high WBGT (Wet Bulb Globe Temperature) or solar radiation, might affect marathon performance under such conditions. Without sufficient data on extreme weather conditions, the analysis may underestimate the true impact of these conditions on performance. Further, the statistical model used in this study does not capture the non-linear relationship between performance and age. More complex models and model selection processes may be needed to address this structure.

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.

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)
# data manipulation
library(dplyr)
library(tidyr)

# summary tables
library(kableExtra)
library(gtsummary)
library(gt)

# interaction plots
library(ggplot2)
library(ggpubr)
library(ggExtra)
library(ggcorrplot)
library(lubridate)
library(ggridges)

# model
# library(lme4)
library(lmerTest)
library(broom.mixed)
# Load datasets
AQI <- read.csv("~/Desktop/PHP 2550/Data/aqi_values.csv")
Course <- read.csv("~/Desktop/PHP 2550/Data/course_record.csv")
Marathon <- read.csv("~/Desktop/PHP 2550/Data/marathon_dates.csv")
Project1<- read.csv("~/Desktop/PHP 2550/Data/project1.csv")

# Modify categorical variables
Project1 <- Project1 %>%
  rename(Race = Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D.,
         Sex = Sex..0.F..1.M.)
Project1$Race <- as.factor(Project1$Race)
Project1$Sex <- as.factor(Project1$Sex)
Project1$Flag[Project1$Flag ==""] <- NA
Project1$Flag <- as.factor(Project1$Flag)
# Create categorical variable age_grp for age and modify humidity varibale

# Distribution of humidity
ggplot(Project1, aes(x = X.rh)) +
  geom_histogram(fill = "skyblue", color="black") +
  labs(x = "Percent Relative Humidity", y = "Count", title = "Figure 1: Histogram of Percent Relative Hu
  theme_minimal()

Project1 <- Project1 %>%
  mutate(age_grp = cut(Age..yr., breaks = c(0, 24, 39, 54, 69, Inf),
                                   include.lowest = TRUE,
                                   labels = c("< 25 yrs", "25-39 yrs", "40-54 yrs", "55-69 yrs", ">= 70 y
         age_grp = factor(age_grp, levels = c("< 25 yrs", "25-39 yrs", "40-54 yrs", "55-69 yrs", ">= 70
         X.rh = ifelse(X.rh <= 1, X.rh*100, X.rh))
```

```r
# Summary Table 1 (N by gender and age group)
Project1 %>%
  select(Sex, age_grp) %>%
  mutate(Sex = ifelse(Sex==0, "Female", "Male")) %>%
  tbl_summary(by = Sex,
              label = list(age_grp ~ "Age Group"),
              statistic = list(
                all_continuous() ~ "{mean} ({sd})",
                all_categorical() ~ "{n}"),
              missing = "ifany",
              missing_text = "Missing") %>%
  modify_spanning_header(update =  all_stat_cols() ~  "**Gender**") %>%
  modify_footnote(update = everything() ~ NA) %>%
  bold_labels() %>%
  as_kable_extra(booktabs = TRUE,
                 caption = "Number of Participants by Age Group and Gender",
                 longtable = TRUE, linesep = "") %>%
  kableExtra::kable_styling(font_size = 10,
                 latex_options = c("repeat_header", "HOLD_position", "scale_down"))
# Summary Table 2
AQI_mod <- AQI %>%
   mutate(Race= case_when(marathon == "Boston" ~ "0",
                          marathon == "Chicago" ~ "1",
                          marathon == "NYC" ~ "2",
                          marathon == "Twin Cities" ~ "3",
                          marathon == "Grandmas" ~ "4",
                          )) %>%
  filter(sample_duration != "1 HOUR") %>%
  group_by(Race, date_local) %>%
  summarize(ave_aqi = mean(aqi, na.rm = TRUE), .groups = "drop") %>%
  mutate(year = year(ymd(date_local))) %>%
  select(Race, ave_aqi, year)

Course_mod <- Course %>%
  mutate(Race= case_when(Race == "B" ~ "0",
                          Race == "C" ~ "1",
                          Race == "NY" ~ "2",
                          Race == "TC" ~ "3",
                          Race == "D" ~ "4",
                          ),
         CR = period_to_seconds(hms(CR)))
# Merge Data
Project1_merged <- Project1 %>%
  left_join(Course_mod, by=c("Race", "Year")) %>%
  left_join(AQI_mod, by = c("Race", "Year" = "year"))

Project1_unique <- Project1_merged %>%
  distinct(Race, Year, .keep_all = TRUE)

Project1_tbl_w <- Project1_unique %>%
  select(Race, Flag, Td..C, Tw..C, X.rh, Tg..C, SR.W.m2, Wind, DP, CR, ave_aqi) %>%
  mutate(
    Race = case_when(Race == 0 ~ "Boston Marathon",
```

```
                    Race == 1 ~ "Chicago Marathon",
                    Race == 2 ~ "New York City Marathon",
                    Race == 3 ~ "Twin Cities Marathon",
                    Race == 4 ~ "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"),
    Flag = factor(Flag, levels = c("WBGT <10C", "WBGT 10-18C", "WBGT >18-23C", "WBGT >23-28C", "Missing
  tbl_summary(by=Race,
              label = list(
                #Age..yr. ~ "Age",
                           #X.CR ~ "Percent off current course record ",
                           Td..C ~ "Dry bulb temperature",
                           Tw..C ~ "Wet bulb temperature",
                           X.rh ~ "Percent relative humidity",
                           Tg..C ~ "Black globe temperature",
                           SR.W.m2 ~ "Solar radiation in Watts",
                           DP ~ "Dew Point",
                           CR ~ "Course Record in seconds",
                           ave_aqi ~ "Average Air Quality"
                           ),
              statistic = list(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()

Project1_tbl_w %>%
  as_kable_extra(booktabs = TRUE,
                 caption = "Summary Statistics of Weather Condistions by Race",
                 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)
# Mean CR% by Age (male)
male_summary <- Project1 %>%
  filter(Sex == 1) %>%
  group_by(Age..yr.) %>%
  summarise(
    mean_CR = mean(X.CR, na.rm = TRUE),
    se_CR = sd(X.CR, na.rm = TRUE)
    # se_CR = sd(X.CR, na.rm = TRUE) / sqrt(n())
  )
```

```r
# Mean CR% by Age (female)
female_summary <- Project1 %>%
  filter(Sex == 0) %>%
  group_by(Age..yr.) %>%
  summarise(
    mean_CR = mean(X.CR, na.rm = TRUE),
    se_CR = sd(X.CR, na.rm = TRUE)
    # se_CR = sd(X.CR, na.rm = TRUE) / sqrt(n())
  )
## 1.1 Distribution in Mean CR% by Age
gender_summary <- merge(male_summary, female_summary, by="Age..yr.") %>%
  pivot_longer(cols = c(mean_CR.x, mean_CR.y), values_to = "mean_CR", names_to = "Sex") %>%
  mutate(age_grp = cut(Age..yr., breaks = c(0, 24, 39, 54, 69, Inf),
                                  include.lowest = TRUE,
                                  labels = c("< 25 yrs", "25-39 yrs", "40-54 yrs", "55-69 yrs", ">= 70 y
         age_grp = factor(age_grp, levels = c("< 25 yrs", "25-39 yrs", "40-54 yrs", "55-69 yrs", ">= 70

# Density Plot for %CR by Age Group
gender_summary %>%
  ggplot(aes(mean_CR, fill = age_grp)) +
  geom_density(alpha = 0.5) +
  labs(x = "Mean %CR (Percent off current course record)",
       y = "Density",
       title = "Figure 2: Distribution in mean %CR by Age Group",
       fill = "Age Group") +
  theme_minimal()
## 1.2 Summary Statistics of %CR by Age Group and Gender
male_summary_grp <- Project1 %>%
  filter(Sex == 1) %>%
  group_by(age_grp) %>%
  summarise(
    mean_CR = mean(X.CR, na.rm = TRUE),  # Compute the mean
    lower = quantile(X.CR, 0.25, na.rm = TRUE),  # Compute the lower quantile (25%)
    upper = quantile(X.CR, 0.75, na.rm = TRUE)  # Compute the upper quantile (75%)
  ) %>%
  mutate(
    Male = paste0(round(mean_CR, 2), " (", round(lower, 2), ", ", round(upper, 2), ")")  # Combine into
  ) %>%
  select(age_grp, Male)

female_summary_grp <- Project1 %>%
  filter(Sex == 0) %>%
  group_by(age_grp) %>%
  summarise(
    mean_CR = mean(X.CR, na.rm = TRUE),  # Compute the mean
    lower = quantile(X.CR, 0.25, na.rm = TRUE),  # Compute the lower quantile (25%)
    upper = quantile(X.CR, 0.75, na.rm = TRUE)  # Compute the upper quantile (75%)
  ) %>%
  mutate(
    Female = paste0(round(mean_CR, 2), " (", round(lower, 2), ", ", round(upper, 2), ")")  # Combine in
  ) %>%
  select(age_grp, Female)
```

```r
# Combine two summary tables
summary_sex_grp <- cbind(male_summary_grp, female_summary_grp)[, -3]
empty_row <- tibble(age_grp = "CR% (by Age Group)", Male = "", Female = "")
summary_sex_grp <- bind_rows(empty_row, summary_sex_grp)

summary_sex_grp %>%
  gt() %>%
  cols_label(
    age_grp = "Characteristic",
    Male = "Men, N=6,112",
    Female = "Women, N=5,452") %>%
  tab_header(
    title = md("Table 3: Summary Statistics of %CR by Gender, Mean (Q1, Q3)")) %>%
  tab_style(
    style = cell_text(size = px(6), weight = "bold"),  # Bold the column labels
    locations = cells_column_labels(columns = c(age_grp, Male, Female))) %>%
  tab_style(
    style = cell_text(size = px(6), weight = "bold"),  # Bold the "CR% (by age group)" row
    locations = cells_body(columns = everything(), rows = age_grp == "CR% by Age Group")) %>%
  tab_style(
    style = cell_text(size = px(8), weight = "bold"),  # Adjust the header size
    locations = cells_title(groups = "title")  # Apply to the title only
  ) %>%
  tab_options(
    table.width = pct(100),
    table.align = "center"
  )
## 1.3 Interaction Plot (Age and %CR by gender)
# Plot (Male)
p_age_male <- ggplot(male_summary, aes(x = Age..yr., y = mean_CR)) +
  geom_point(color = "grey", size = 1) +          # Add points for the mean
  geom_errorbar(aes(ymin = mean_CR - se_CR, ymax = mean_CR + se_CR), width = 1, color = "grey") +  # Ad
  geom_smooth(se = FALSE, color = "black", size = 1, method = "loess", linetype = 2) +  # Add smooth li
  geom_vline(xintercept = 70, linetype = "dashed", color = "red", size = 0.8) +
  annotate("text", x = 65, y = 300, label = "Age = 70", color = "red", angle = 90, size = 4, hjust = 0)
  geom_point(aes(x = 30, y = 7.13), color = "blue", size = 2) +
  geom_hline(yintercept = 7.13, linetype = "dashed", color = "blue", size = 0.8) +
  annotate("text", x = 25, y = 60, label = "(30, 7.13)", color = "blue", angle = 0, size = 4, hjust = 0
  labs(title = "Men", x = "Age (yrs)", y = "Best Time (%CR)") +
  ylim(0, 400) +
  theme_minimal(base_size = 15) +
  theme(plot.title = element_text(hjust = 0.5, size=14),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12))

# Plot (Female)
p_age_female <- ggplot(female_summary, aes(x = Age..yr., y = mean_CR)) +
  geom_point(color = "grey", size = 1) +          # Add points for the mean
  geom_errorbar(aes(ymin = mean_CR - se_CR, ymax = mean_CR + se_CR), width = 1, color = "grey") +  # Ad
  geom_smooth(se = FALSE, color = "black", size = 1, method = "loess", linetype = 2) +  # Add smooth li
  geom_vline(xintercept = 70, linetype = "dashed", color = "red", size = 0.8) +
  annotate("text", x = 65, y = 300, label = "Age = 70", color = "red", angle = 90, size = 4, hjust = 0)
  geom_point(aes(x = 28, y = 12.27), color = "blue", size = 2) +
```

```r
  geom_hline(yintercept = 12.27, linetype = "dashed", color = "blue", size = 0.8) +
  annotate("text", x = 25, y = 60, label = "(28, 12.27)", color = "blue", angle = 0, size = 4, hjust = 
  labs(title = "Women", x = "Age (yrs)", y = "Best Time (%CR)") +
  ylim(0, 400) +
  theme_minimal(base_size = 15) +
  theme(plot.title = element_text(hjust = 0.5, size=14),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12))

# Merge
ggarrange(p_age_male+ rremove("ylab") + rremove("xlab"),
          p_age_female+ rremove("ylab") + rremove("xlab"),
          ncol = 2, nrow = 1,
          labels = c("A", "B"),
          font.label = list(size = 12),
          common.legend = TRUE, align = "hv") %>%
  annotate_figure(
    top = text_grob("Figure 3: Interaction between %CR and Age by Gender",
                    face = "bold", size = 14),
    left = text_grob("Best Time (%CR)", rot = 90),
    bottom = text_grob("Age (yrs)") )
## 2.1 WBGT (Temperature)
women_summary_WBGT <- Project1 %>%
  filter(Sex == 0) %>%
  group_by(WBGT, age_grp) %>%
  summarise(
    mean_CR = mean(X.CR, na.rm = TRUE)
    # ,
    # se_CR = sd(X.CR, na.rm = TRUE)
  )

men_summary_WBGT <- Project1 %>%
  filter(Sex == 1) %>%
  group_by(WBGT, age_grp) %>%
  summarise(
    mean_CR = mean(X.CR, na.rm = TRUE)
    # ,
    # se_CR = sd(X.CR, na.rm = TRUE)
  )

p_WBGT_women <- ggplot(women_summary_WBGT, aes(x = WBGT, y = mean_CR, color = age_grp)) +
  geom_point(size = 0.8, alpha = 0.5) +          # Add points for the mean
  geom_smooth(se = TRUE, size = 1, method = "loess", linetype = 2) +  # Add smooth line without confide
  # facet_zoom(ylim = c(100, 150)) +
  labs(title = "Women", x = "WBGT", y = "Best Time (%CR)", color = "Age Group") +  # Add labels and tit
  ylim(0, 200) +                                 # Set y-axis limit
  theme_minimal(base_size = 15) +                # Use minimal theme with larger font
  theme(plot.title = element_text(hjust = 0.5, size = 14),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12),
        legend.text = element_text(size = 10),
        legend.title = element_text(size = 12))
```

```r
p_WBGT_men <- ggplot(men_summary_WBGT, aes(x = WBGT, y = mean_CR, color = age_grp)) +
  geom_point(size = 0.8, alpha = 0.5) +          # Add points for the mean
  geom_smooth(se = TRUE, size = 1, method = "loess", linetype = 2) +  # Add smooth line without confide
  labs(title = "Men", x = "WBGT", y = "Best Time (%CR)", color = "Age Group") +  # Add labels and title
  ylim(0, 200) +                                  # Set y-axis limit
  theme_minimal(base_size = 15) +                 # Use minimal theme with larger font
  theme(plot.title = element_text(hjust = 0.5, size = 14),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12),
        legend.text = element_text(size = 10),
        legend.title = element_text(size = 12))

ggarrange(p_WBGT_women+ rremove("ylab"),
                    p_WBGT_men+ rremove("ylab"),
                    common.legend = TRUE, legend = "bottom", align = "hv") %>%
  annotate_figure(top = text_grob("Figure 4: Interaction between %CR and WBGT with Age by Gender", face
                left = text_grob("Best Time (%CR)", rot = 90, size = 15))

women_summary_hum <- Project1 %>%
  filter(Sex == 0) %>%
  group_by(X.rh, age_grp) %>%
  summarise(
    mean_CR = mean(X.CR, na.rm = TRUE)
    # ,
    # se_rh = sd(X.rh, na.rm = TRUE)
  )

men_summary_hum <- Project1 %>%
  filter(Sex == 1) %>%
  group_by(X.rh, age_grp) %>%
  summarise(
    mean_CR = mean(X.CR, na.rm = TRUE)
    # ,
    # se_rh = sd(X.rh, na.rm = TRUE)
  )

p_hum_women <- ggplot(women_summary_hum, aes(x = X.rh, y = mean_CR, color = age_grp)) +
  geom_point(size = 0.8, alpha = 0.5) +          # Add points for the mean
  geom_smooth(se = TRUE, size = 1, method = "loess", linetype = 2) +  # Add smooth line without confide
  labs(title = "Women", x = "Percent relative humidity", y = "Best Time (%CR)", color = "Age Group") +
  ylim(0, 200) +                                  # Set y-axis limit
  theme_minimal(base_size = 15) +                 # Use minimal theme with larger font
  theme(plot.title = element_text(hjust = 0.5, size = 14),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12),
        legend.text = element_text(size = 10),
        legend.title = element_text(size = 12))

p_hum_men <- ggplot(men_summary_hum, aes(x = X.rh, y = mean_CR, color = age_grp)) +
  geom_point(size = 0.8, alpha = 0.5) +        # Add points for the mean
  geom_smooth(se = TRUE, size = 1, method = "loess", linetype = 2) +  # Add smooth line without confide
  labs(title = "Men", x = "Percent relative humidity", y = "Best Time (%CR)", color = "Age Group") +  #
  ylim(0, 200) +                                  # Set y-axis limit
```

```r
  theme_minimal(base_size = 15) +                  # Use minimal theme with larger font
  theme(plot.title = element_text(hjust = 0.5, size = 14),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12),
        legend.text = element_text(size = 10),
        legend.title = element_text(size = 12))

ggarrange(p_hum_women+ rremove("ylab"),
                    p_hum_men+ rremove("ylab"),
                    common.legend = TRUE, legend = "bottom", align = "hv") %>%
  annotate_figure(top = text_grob("Figure 5: Interaction between %CR and Humidity with Age by Gender", 
    left = text_grob("Best Time (%CR)", rot = 90, size = 15))
women_summary_SR <- Project1 %>%
  filter(Sex == 0) %>%
  group_by(SR.W.m2, age_grp) %>%
  summarise(
    mean_CR = mean(X.CR, na.rm = TRUE),
    se_CR = sd(X.CR, na.rm = TRUE)
  )

men_summary_SR <- Project1 %>%
  filter(Sex == 1) %>%
  group_by(SR.W.m2, age_grp) %>%
  summarise(
    mean_CR = mean(X.CR, na.rm = TRUE),
    se_CR = sd(X.CR, na.rm = TRUE)
  )

p_SR_women <- ggplot(women_summary_SR, aes(x = SR.W.m2, y = mean_CR, color = age_grp)) +
  geom_point(size = 0.8, alpha = 0.5) +          # Add points for the mean
  geom_smooth(se = TRUE, size = 0.8, method = "loess", linetype = 2) +  # Add smooth line without confi
  labs(title = "Women", x = "Solar Radiatio", y = "Best Time (%CR)", color = "Age Group") +  # Add labe
  ylim(0, 200) +                                   # Set y-axis limit
  theme_minimal(base_size = 15) +                  # Use minimal theme with larger font
  theme(plot.title = element_text(hjust = 0.5, size = 14),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12),
        legend.text = element_text(size = 10),
        legend.title = element_text(size = 12))

p_SR_men <- ggplot(men_summary_SR, aes(x = SR.W.m2, y = mean_CR, color = age_grp)) +
  geom_point(size = 0.8, alpha = 0.5) +         # Add points for the mean
  geom_smooth(se = TRUE, size = 0.8, method = "loess", linetype = 2) +  # Add smooth line without confi
  labs(title = "Men", x = "Solar Radiation", y = "Best Time (%CR)", color = "Age Group") +  # Add label
  ylim(0, 200) +                                   # Set y-axis limit
  theme_minimal(base_size = 15) +                  # Use minimal theme with larger font
  theme(plot.title = element_text(hjust = 0.5, size = 14),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12),
        legend.text = element_text(size = 10),
        legend.title = element_text(size = 12))

ggarrange(p_SR_women+ rremove("ylab"),
```

```
                    p_SR_men+ rremove("ylab"),
                    common.legend = TRUE, legend = "bottom", align = "hv") %>%
  annotate_figure(top = text_grob("Figure 6: Interaction between %CR and Solar Radiation with Age by Ge
                  left = text_grob("Best Time (%CR)", rot = 90, size = 15))
women_summary_wind <- Project1 %>%
  filter(Sex == 0) %>%
  group_by(Wind, age_grp) %>%
  summarise(
    mean_CR = mean(X.CR, na.rm = TRUE),
    se_CR = sd(X.CR, na.rm = TRUE)
  )

men_summary_wind <- Project1 %>%
  filter(Sex == 1) %>%
  group_by(Wind, age_grp) %>%
  summarise(
    mean_CR = mean(X.CR, na.rm = TRUE),
    se_CR = sd(X.CR, na.rm = TRUE)
  )

p_wind_women <- ggplot(women_summary_wind, aes(x = Wind, y = mean_CR, color = age_grp)) +
  geom_point(size = 0.8, alpha = 0.5) +        # Add points for the mean
  geom_smooth(se = TRUE, size = 0.8, method = "loess", linetype = 2) +  # Add smooth line without confi
  labs(title = "Women", x = "Wind", y = "Best Time (%CR)", color = "Age Group") +  # Add labels and tit
  ylim(0, 200) +                                    # Set y-axis limit
  theme_minimal(base_size = 15) +                   # Use minimal theme with larger font
  theme(plot.title = element_text(hjust = 0.5, size = 14),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12),
        legend.text = element_text(size = 10),
        legend.title = element_text(size = 12))

p_wind_men <- ggplot(men_summary_wind, aes(x = Wind, y = mean_CR, color = age_grp)) +
  geom_point(size = 0.8, alpha = 0.5) +         # Add points for the mean
  geom_smooth(se = TRUE, size = 0.8, method = "loess", linetype = 2) +  # Add smooth line without confi
  labs(title = "Men", x = "Wind", y = "Best Time (%CR)", color = "Age Group") +  # Add labels and title
  ylim(0, 200) +                                    # Set y-axis limit
  theme_minimal(base_size = 15) +                   # Use minimal theme with larger font
  theme(plot.title = element_text(hjust = 0.5, size = 14),
        axis.title.x = element_text(size = 12),
        axis.title.y = element_text(size = 12),
        legend.text = element_text(size = 10),
        legend.title = element_text(size = 12))

ggarrange(p_wind_women+ rremove("ylab"),
                    p_wind_men+ rremove("ylab"),
                    common.legend = TRUE, legend = "bottom", align = "hv") %>%
  annotate_figure(top = text_grob("Figure 7: Interaction between %CR and Wind with Age by Gender", face
    left = text_grob("Best Time (%CR)", rot = 90, size = 15))
Project1_name <- Project1_merged %>%
  rename("% CR" = X.CR,
         "Dry bulb Temp" = Td..C,
         "Wet bulb Temp" = Tw..C,
```

```
         "Percent relative humidity" = X.rh,
         "Black globe Temp" = Tg..C,
         "Solar Radiation" = SR.W.m2,
         "Average AQI" = ave_aqi,
         "Dew Point" = DP
         )
r <- cor(Project1_name[, c(6:14,17)], use="complete.obs")

variable_order <- c("Dry bulb Temp", "Wet bulb Temp",
                    "Black globe Temp", "Solar Radiation",
                    "Percent relative humidity", "Average AQI", "Dew Point", "WBGT", "% CR")

# Reorder the correlation matrix based on the new variable order
r_reordered <- r[variable_order, variable_order]

ggcorrplot(r_reordered,
           hc.order = TRUE,
           type = "lower",
           lab = TRUE) +
  ggtitle("Figure 8: Correlation Matrix of Environmental Variables") +
  theme(plot.title = element_text(hjust = 0.5, size=14),
        axis.text.x = element_text(size = 12),
        axis.text.y = element_text(size = 12))
# model
M1 <- lmerTest::lmer(X.CR ~ WBGT + X.rh + SR.W.m2 + Wind + ave_aqi + (1|Age..yr.),
                     data=Project1_merged)

tidy_m1_fixed <- tidy(M1, effects = "fixed")
tidy_m1_random <- tidy(M1, effects = "ran_pars")

tidy_m1_fixed$term <- c("(Intercept)", "WBGT", "Percent Humidity",
                        "Solar Radiation", "Wind", "Average AQI")
tidy_m1_random$group <- c("Age", "Residual")
tidy_m1_random$term <- c("(Intercept)", "")

tidy_m1_fixed[, -c(1, 5, 6)] %>%
  kbl(booktabs = TRUE, caption = "Fixed Effects of the Model",
      col.names = c("Term", "Estimate", "Standard Error", "P-value"),
      longtable = TRUE, linesep = "") %>%
  kable_styling(font_size = 10,
                latex_options = c("repeat_header", "HOLD_position", "scale_down"))

# Print random effects using kable
tidy_m1_random[1,3] <- "Intercept"
tidy_m1_random[, -1] %>%
  kbl(booktabs = TRUE, caption = "Random Effect of the Model",
      col.names = c("Group", "Term","SD"),
      longtable = TRUE, linesep = "") %>%
  kable_styling(font_size = 10,
                latex_options = c("repeat_header", "HOLD_position", "scale_down"))
```