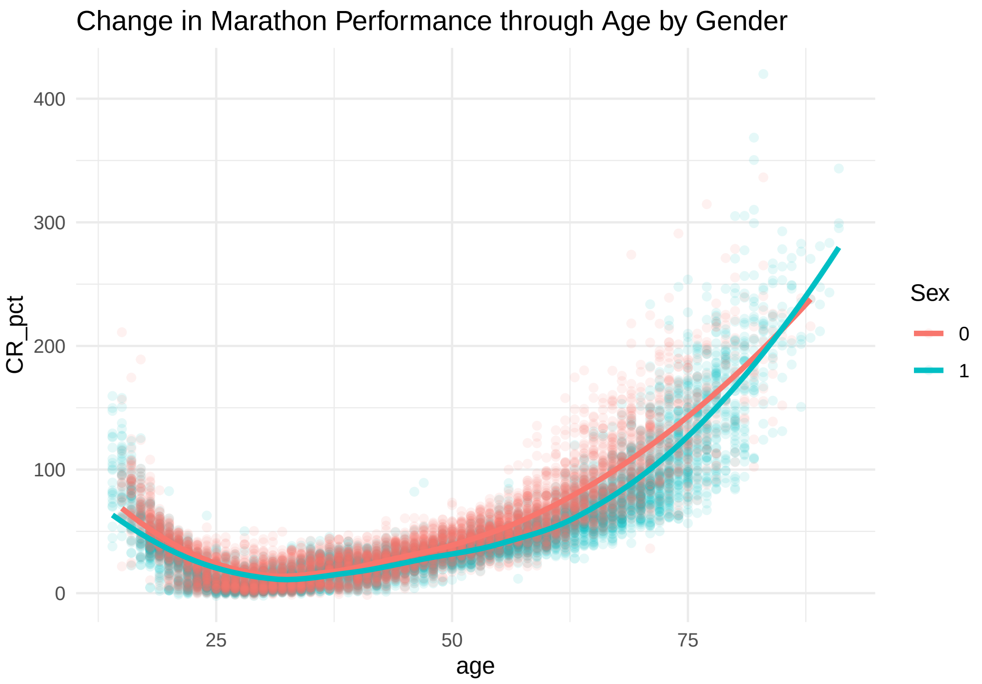
Project1 Codebook

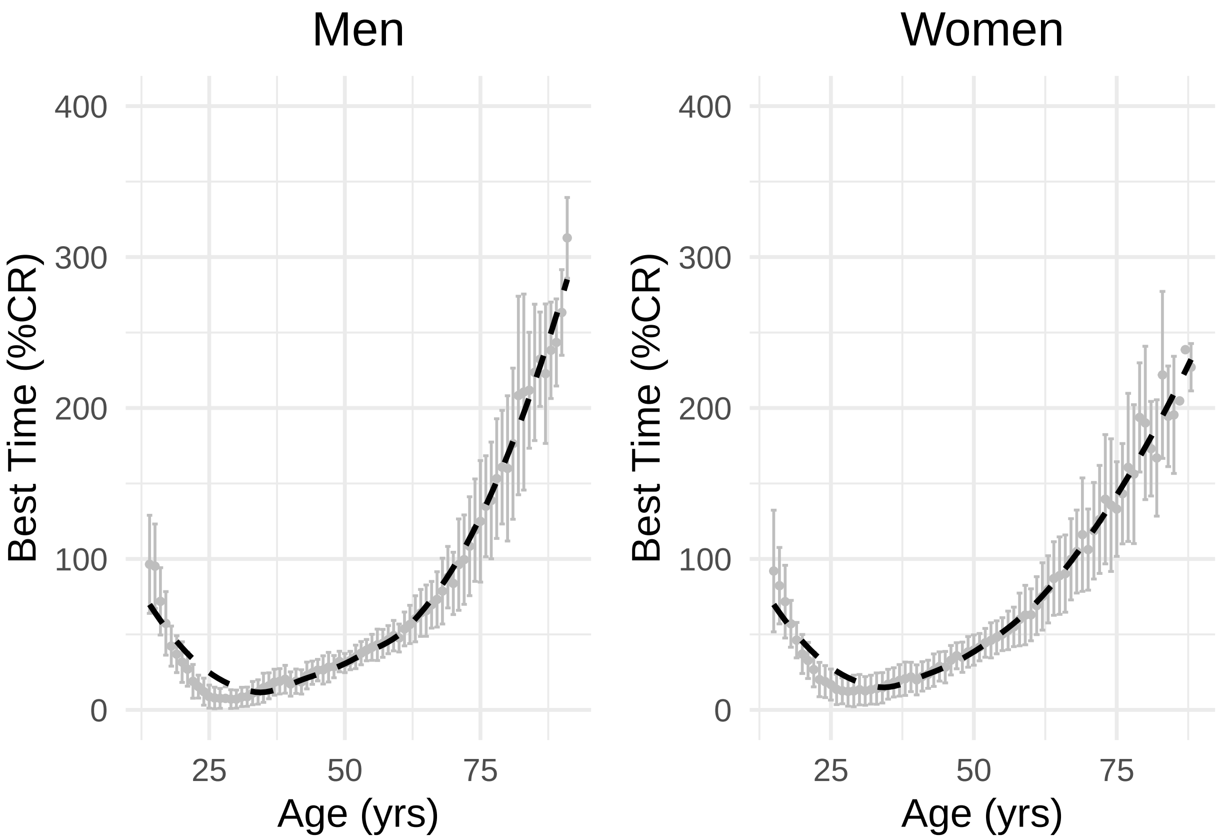
Yingxi Kong

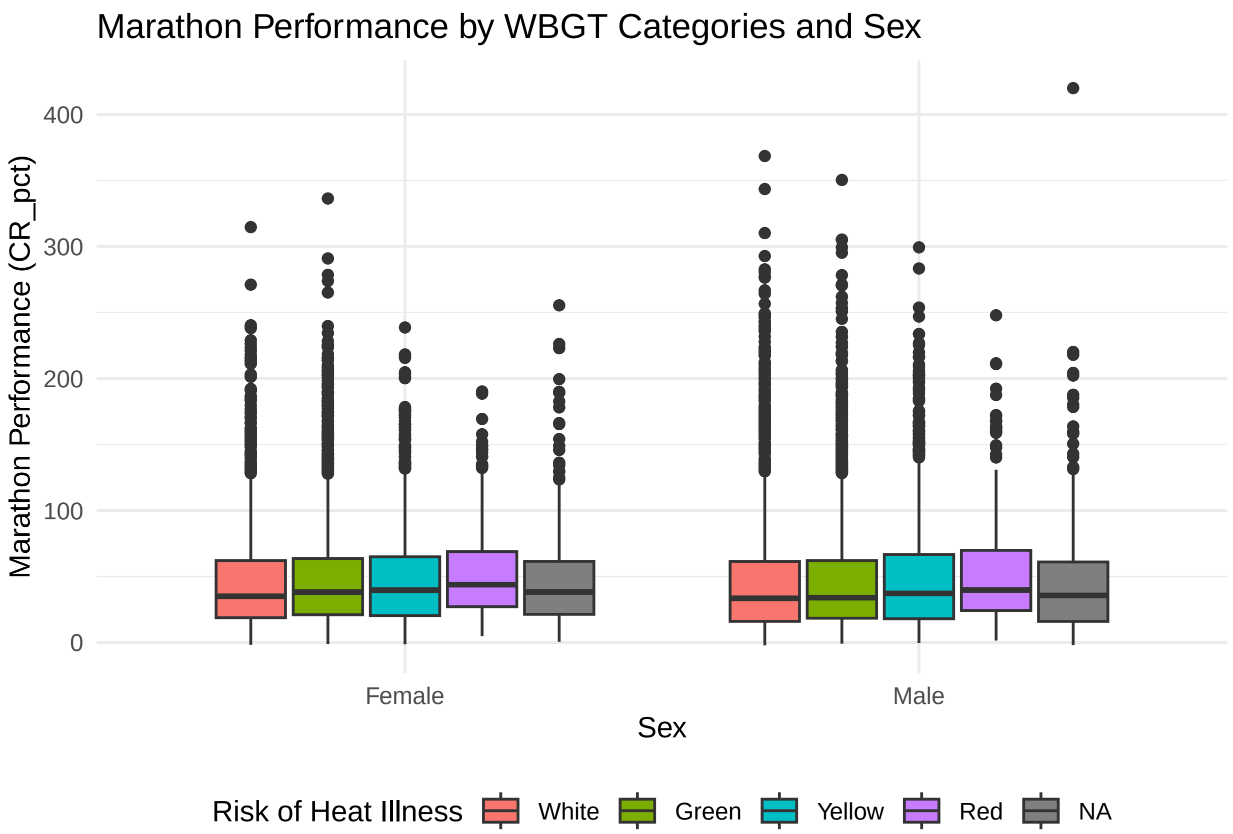
2024-10-01

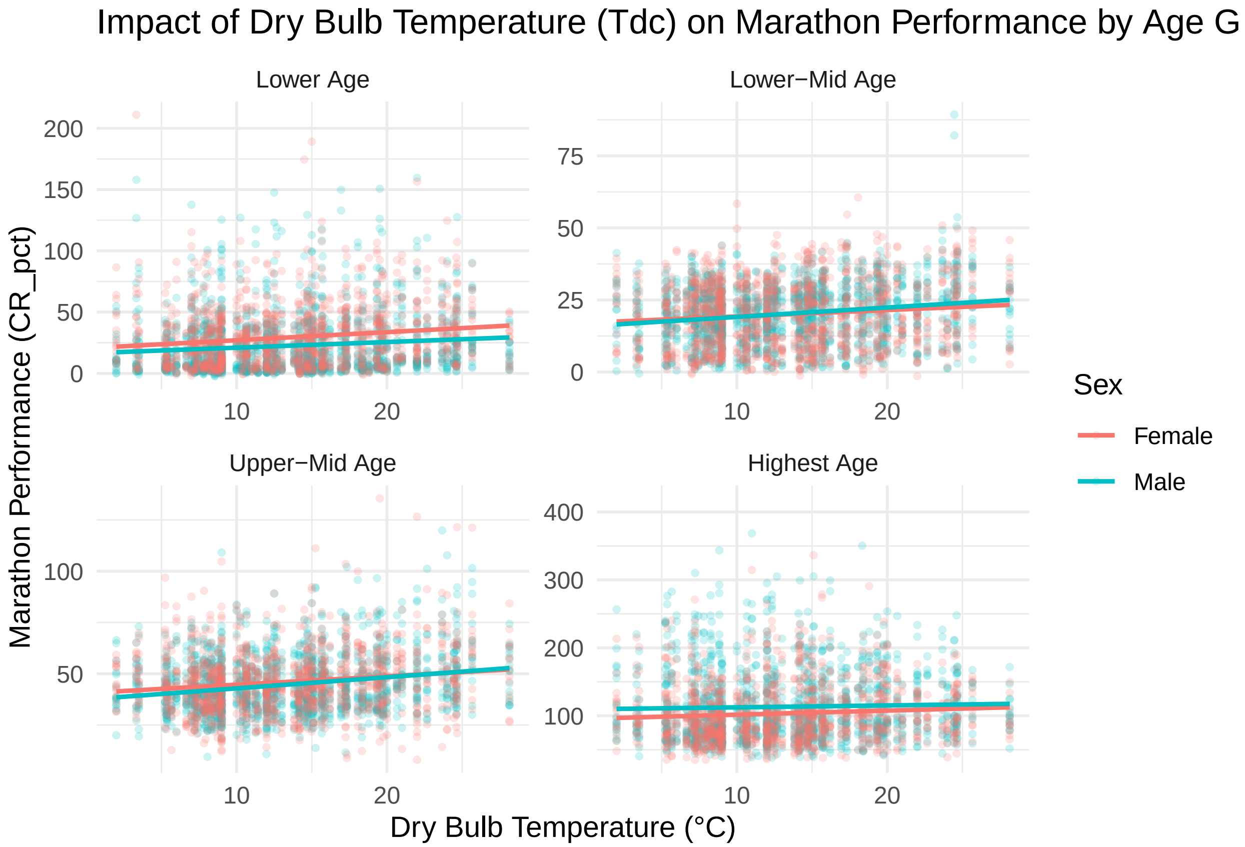
Table 1: Summary of Marathon Performance by Age Group and Sex

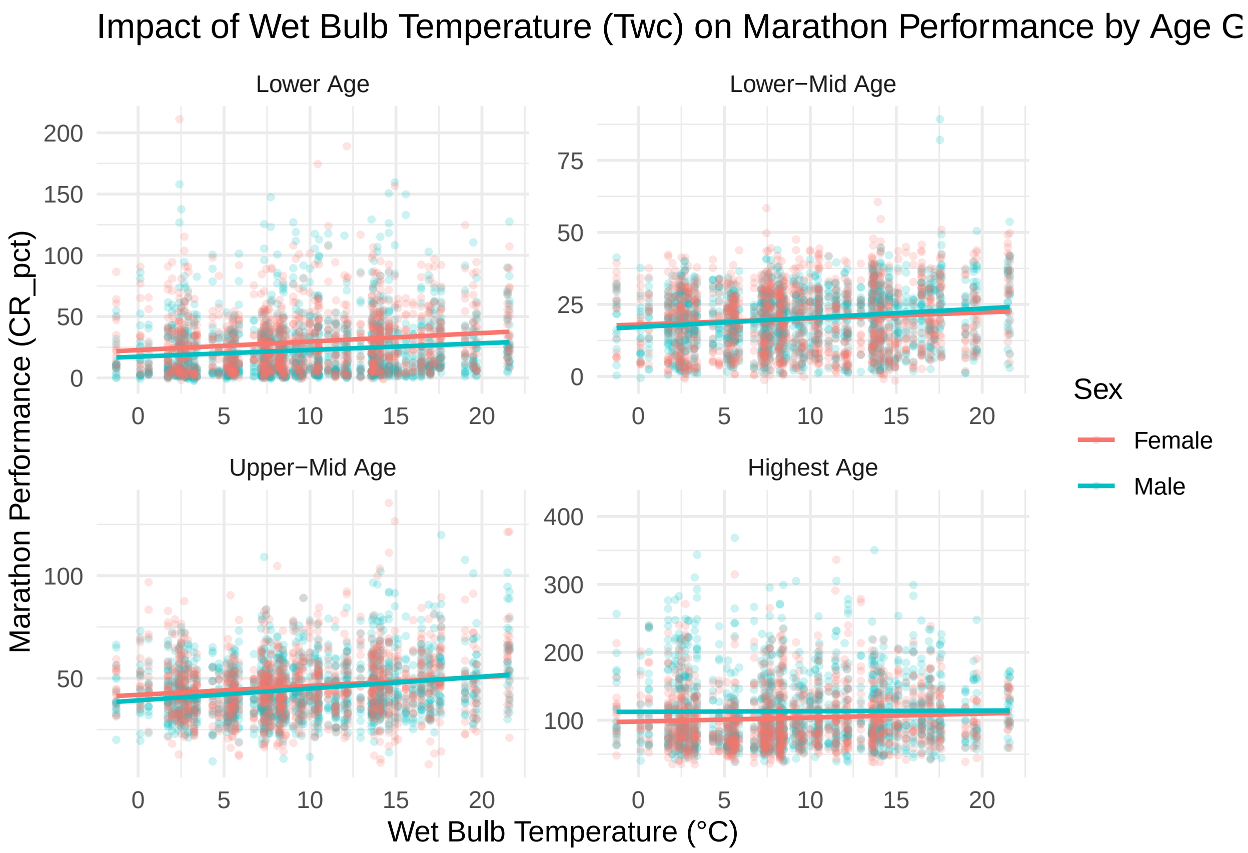
Age Group	Sex	N	Min Performance	Mean Performance	Median Performance	Max Performance
Lower Age	Female	1364	-1.816	29.310	23.870	211.095
Lower Age	Male	1602	-2.251	22.609	13.748	159.535
Lower-Mid Age	Female	1440	-1.419	20.070	19.973	60.567
Lower-Mid Age	Male	1536	-0.499	20.189	20.490	89.271
Upper-Mid Age	Female	1343	8.045	45.882	44.877	135.478
Upper-Mid Age	Male	1535	9.310	44.677	41.996	119.853
Highest Age	Female	1305	35.119	103.639	94.332	336.347
Highest Age	Male	1439	38.345	113.462	97.587	419.958

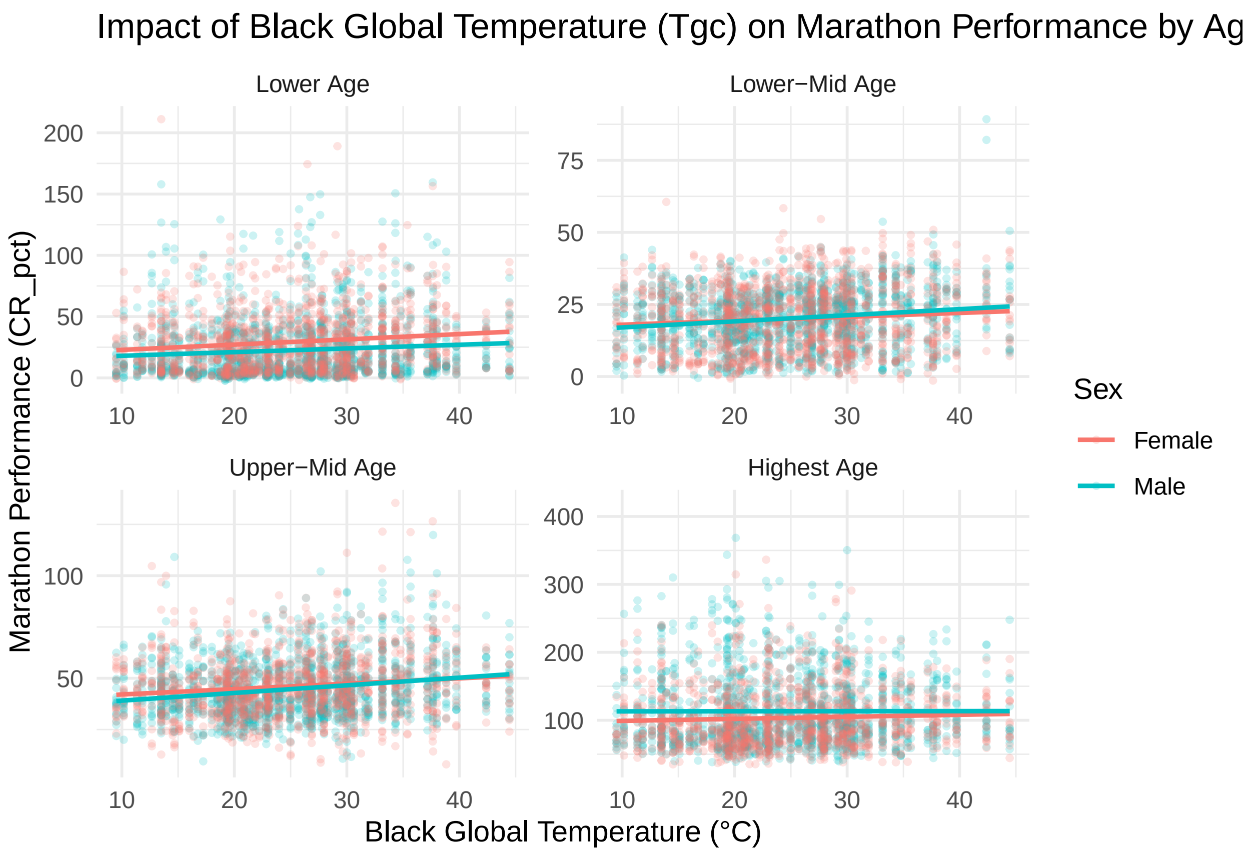



Men
Women
Best Time (%CR)
400
300
200
100
0
25
50
75
Age (yrs)


Marathon Performance by WBGT Categories and Sex
Marathon Performance (CR_pct)
400
300
200
100
0
Female
Male
Sex
Risk of Heat Illness
White
Green
Yellow
Red
NA


Impact of Dry Bulb Temperature (Tdc) on Marathon Performance by Age G
Lower Age
Lower−Mid Age
Upper−Mid Age
Highest Age
Marathon Performance (CR_pct)
Dry Bulb Temperature (°C)
0
50
100
150
200
0
25
50
75
50
100
100
200
300
400
10
20
Sex
Female
Male


Impact of Wet Bulb Temperature (Twc) on Marathon Performance by Age G
Lower Age
Lower−Mid Age
Upper−Mid Age
Highest Age
Marathon Performance (CR_pct)
Wet Bulb Temperature (°C)
0
5
10
15
20
200
150
100
50
0
75
50
25
0
100
50
400
300
200
100
Sex
Female
Male



Appendix

```
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, message = FALSE)
library(tidyverse)
library(ggplot2)
library(visdat)
library(gtsummary)
library(kableExtra)
library(ggpubr)

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")

colnames(data) <- c("race", "year", "sex", "flag", "age", "CR_pct", "Tdc", "Twc", "rh", "Tgc", "SRWm2",

data$flag <- case_when(data$flag == "" ~ NA, TRUE ~ data$flag)
data$flag <- as.factor(data$flag)

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,
```

```r
                                course_record$Race == "D" ~ 4,
                                TRUE ~ NA)
course_record$Gender <- case_when(course_record$Gender == "M" ~ 1,
                                  course_record$Gender == "F" ~ 0,
                                  TRUE ~ NA)

data <- data %>%
  left_join(course_record, join_by("sex" == "Gender", "race" == "Race", "year" == "Year"))

data$sex <- as.factor(data$sex)
data$race <- as.factor(data$race)
# Missing Pattern
# vis_dat(data)

# Classify each observation to age groups by gender's quantile value
data <- data %>%
  group_by(sex) %>%
  mutate(age_group = cut(age, breaks = quantile(age, probs = seq(0, 1, 0.25), na.rm = TRUE),
                                  include.lowest = TRUE,
                                  labels = c("Lower Age", "Lower-Mid Age",
                                             "Upper-Mid Age", "Highest Age")))
# 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))
summary_table$sex <- ifelse(summary_table$sex == 1, "Male", "Female")

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)


ggplot(data) +
  geom_point(aes(x = age, y = CR_pct, color = sex), alpha = 0.1) +
  geom_smooth(aes(x = age, y = CR_pct, color = sex), method = "loess", se = FALSE, size = 1.2) +
  ggtitle("Change in Marathon Performance through Age by Gender") +
  theme_minimal() +
  labs(color = "Sex")

# male summary
male_summary <- data %>%
  filter(sex == 1) %>%
  group_by(age) %>%
  summarise(mean_CR = mean(CR_pct, na.rm = TRUE),
            se_CR = sd(CR_pct, na.rm = TRUE))
```

```r
# create the plot
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) +
  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))

# women summary
women_summary <- data %>%
  filter(sex == 0) %>%
  group_by(age) %>%
  summarise(mean_CR = mean(CR_pct, na.rm = TRUE),
            se_CR = sd(CR_pct, na.rm = TRUE))

# create the plot
ageplot_female <- ggplot(women_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) +
  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))

# merge the two plots together
ggarrange(ageplot_male, ageplot_female)
data$flag <- factor( data$flag, levels = c("White", "Green", "Yellow", "Red", "Black", NA))
data$sex <- ifelse(data$sex == 0, "Female", "Male")
ggplot(data, aes(x = sex, y = CR_pct, fill = flag)) +
  geom_boxplot() +
  ggtitle("Marathon Performance by WBGT Categories and Sex") +
  theme_minimal() +
  labs(x = "Sex", y = "Marathon Performance (CR_pct)", fill = "Risk of Heat Illness") +
  theme(legend.position = "bottom")
ggplot(data, aes(x = Tdc, y = CR_pct, color = sex)) +
  geom_point(alpha = 0.2, size = 0.8) +
  geom_smooth(method = "lm", se = FALSE, size = 0.8) +
  facet_wrap(~ age_group, scales = "free") +
  ggtitle("Impact of Dry Bulb Temperature (Tdc) on Marathon Performance by Age Group and Sex") +
  theme_minimal() +
  labs(x = "Dry Bulb Temperature (°C)", y = "Marathon Performance (CR_pct)", color = "Sex")

ggplot(data, aes(x = Twc, y = CR_pct, color = sex)) +
  geom_point(alpha = 0.2, size = 0.8) +
  geom_smooth(method = "lm", se = FALSE, size = 0.8) +
  facet_wrap(~ age_group, scales = "free") +
  ggtitle("Impact of Wet Bulb Temperature (Twc) on Marathon Performance by Age Group and Sex") +
  theme_minimal() +
  labs(x = "Wet Bulb Temperature (°C)", y = "Marathon Performance (CR_pct)", color = "Sex")
```

```r
ggplot(data, aes(x = Tgc, y = CR_pct, color = sex)) +
  geom_point(alpha = 0.2, size = 0.8) +
  geom_smooth(method = "lm", se = FALSE, size = 0.8) +
  facet_wrap(~ age_group, scales = "free") +
  ggtitle("Impact of Black Global Temperature (Tgc) on Marathon Performance by Age Group and Sex") +
  theme_minimal() +
  labs(x = "Black Global Temperature (°C)", y = "Marathon Performance (CR_pct)", color = "Sex")
```