Project1_Jizhou

Jizhou Tian

2024-10-01

Introduction

Endurance exercise performance is well-known to be affected by environmental conditions. Previous research has shown that rising environmental temperatures negatively impact the performance, with more significant effects in longer-distance events such as marathons (Ely et al., 2007). The extent of this impact may vary depending on both sex and age. In this study, we analyzed data from five major marathon races to investigate the effects of weather on marathon performance across the lifespan in both men and women. We conducted an exploratory analysis aimed at three key objectives: examining how increasing age affects marathon performance in both men and women, exploring how environmental conditions influence performance and whether these effects vary by age and gender, and identifying which weather factors have the greatest impact on marathon outcomes.

Methods

Data Description

We will use marathon data, course record data, and AQI data in the subsequent analysis. The marathon data contains 11564 observations and 14 variables. It includes information from five major marathon races — Boston, Chicago, New York City, Twin Cities, and Grandma's Marathon — collected over different time periods. The earliest records begin in 1993, and the latest end in 2016, but the periods vary across the races. The gender, age and the percent off current course record for each gender-age group are collected. Weather variables related to temperature, humidity, solar radiation, and wind speed are documented. The Wet Bulb Globe Temperature (WBGT) is calculated using dry bulb, wet bulb, and black globe temperature, following the formula below:

$$WBGT = (0.7 * Tw) + (0.2 * Tg) + (0.1 * Td)$$

The course record data includes the race name, year, gender group, and the corresponding course record (time in seconds). The AQI data provides information about the Air Quality Index (AQI) and specific air pollutants recorded at the marathon site during the race.

An overview of marathon data

Variable	Description
Race	0 = Boston Marathon, 1 = Chicago Marathon, 2 = New York City Marathon, 3 = Twin Cities Marathon (Minneapolis MN), 4 = Grandma's Marathon (Duluth MN)
Year	Year of the marathon race
Sex	0 = Female, 1 = Male

Variable	Description
Flag	White = WBGT <10°C, Green = WBGT 10-18°C, Yellow = WBGT >18-23°C, Red = WBGT >23-28°C, Black = WBGT >28°C
Age	Age of marathon participants
CR	Fastest finishing time among men and women at each year of age, compared with the course record
Td	Dry bulb temperature in Celsius
Tw	Wet bulb temperature in Celsius
rh	Percent relative humidity
Tg	Black globe temperature in Celsius
SR	Solar radiation in Watts per meter squared
DP	Dew Point in Celsius
Wind	Wind speed in Km/hr
WBGT	Wet Bulb Globe Temperature (Weighted average of dry bulb, wet bulb, and globe temperature)

Data Processing

As we focus on marathon performance in further analysis, it's important to first define "performance." The variable "CR" in the marathon data represents the percentage by which the fastest finishing time for each gender-age group differs from the current course record. It's important to note that the course record is a non-increasing trend over time. For example, being 5% off the course record in 1996 does not indicate the same speed level as being 5% off in 2016, as the record may have improved over those 20 years. Therefore, to make the results comparable, we calculate the best time (in hours) for each gender-age group in a given year and race. This is done using the variable "CR" and the absolute course record (in seconds), with the following formula:

$$Best\ time = (1 + CR) * Absolute\ course\ record\ /\ 3600$$

Since multiple sites in each marathon city collect air pollution data, we use the mean AQI recorded on the race day for each marathon. We chose not to include other pollutants due to missing values across the years 1993 to 2016. As a comprehensive index, we believe AQI is more representative.

We create binary variables for AQI and wind speed separately. The AQI ranges from 14 to 86. If AQI is below 50, it indicates good air quality, while values between 50 and 100 indicate moderate air quality. For wind speed, which ranges from 0 to 22 km/hour, we used 10 km/hour as the threshold.

Results

Missing Data

We first check whether the five marathon races were recorded continuously from their start until 2016. The results are shown in the table. We find that Boston is missing information for the year 1999, and New York is missing information for the year 2012.

We then check for missing values and find that there are no missing values for AQI or our outcome of interest—the best time for each gender-age group. However, 491 observations are missing all weather condition variables, including WBGT, and these variables are missing simultaneously. The missing data occur in the Chicago, New York City, and Twin Cities marathons in 2011, and in Grandma's marathon in 2012. In these four races, the best times for each gender-age group are recorded, but all weather condition variables are missing. We also notice that the variable "flag" has no "black" values, meaning no race was recorded with WBGT above 28°C. We cannot determine if the lack of WBGT > 28°C values is related to the missing weather condition data, because it is possible that WBGT remained below 28°C across all races.

Table 2: Marathon Race Frequency and Year Range Overview

Race	Number of Times	Min Year	Max Year
Boston	18	1998	2016
Chicago	21	1996	2016
NYC	23	1993	2016
Twin Cities	17	2000	2016
Grandma	17	2000	2016

Since the missing rate (491/10451 = 4.7%) is small, the missing data can be considered negligible. We then drop all the missing cases and proceed with further analysis based on complete cases. After this step, we now have 11073 observations.

Distribution of age and sex across races

Since our data comes from five different races, we want to check if the distribution of sex and age is balanced among them. For convenience, we divide age into groups. As shown in the table, the p-value of 0.9 for gender indicates that there is no statistically significant difference in the proportion of males and females across the different marathon races. The p-value less than 0.001 for age group indicates that there is a statistically significant difference in the distribution of age groups across the different marathon races. However, since the majority of participants (>80%) in each city race are between 20 and 70 years old, the age distribution within this range is similar across races. This suggests that the sample size and performance may vary for both very young and very old participants.

Table 3: **Distribution of Sex and Age across Races**

Variables	**Boston** N = 2,088	**Chicago** N = 2,427	**Grandma** N = 1,884	**NYC** N = 2,799	**Twin Cities** N = 1,875	**p-value**
Gender						0.9
Female	984 (47%)	1,150 (47%)	880 (47%)	1,337 (48%)	867 (46%)	
Male	1,104 (53%)	1,277 (53%)	1,004 (53%)	1,462 (52%)	1,008 (54%)	
Age Group						<0.001
0-19	67 (3.2%)	171 (7.0%)	167 (8.9%)	88 (3.1%)	140 (7.5%)	
20-29	360 (17%)	400 (16%)	320 (17%)	440 (16%)	319 (17%)	
30-39	360 (17%)	400 (16%)	320 (17%)	440 (16%)	320 (17%)	
40-49	360 (17%)	400 (16%)	320 (17%)	440 (16%)	320 (17%)	
50-59	359 (17%)	400 (16%)	320 (17%)	440 (16%)	319 (17%)	
60-69	337 (16%)	391 (16%)	286 (15%)	438 (16%)	294 (16%)	
70-79	215 (10%)	237 (9.8%)	134 (7.1%)	378 (14%)	146 (7.8%)	
80-100	30 (1.4%)	28 (1.2%)	17 (0.9%)	135 (4.8%)	17 (0.9%)	

Examine effects of increasing age on marathon performance in men and women

As shown in the figure, we fit a smooth line for each gender, with the X-axis representing age in years and the Y-axis representing the best time in hours. The figure reveals a clear trend where marathon performance improves, reflected by a decrease in best time, up until the mid-20s for both men and women. After reaching a peak around this age, performance gradually declines as age increases.

For both men and women, the best times are observed between the ages of approximately 20 and 30. As age increases beyond 30, the rate of decline in performance becomes more significant, particularly after age 50, where we observe a sharper upward trend in best finishing time.

Notably, the increase in marathon best time becomes significantly steeper after age 65 for both genders, suggesting that aging has a more substantial impact on performance in older adults. However, the overall pattern of decline is similar between men and women, although women tend to have slightly slower times than men at old ages.

The consistent gap between male and female performance across all age groups highlights gender differences, though the overall trend of performance decline with increasing age is shared by both genders.

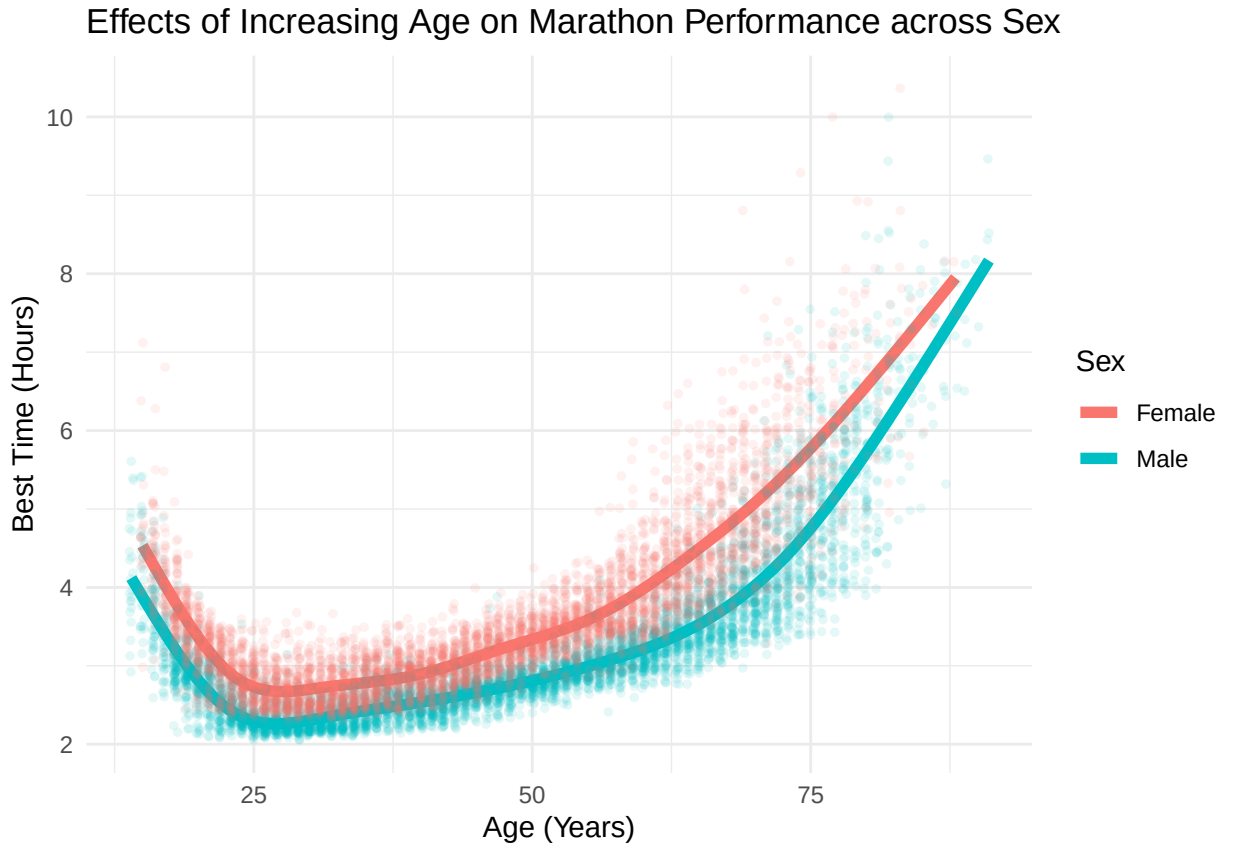

Figure 1: Effects of Increasing Age on Marathon Performance across Sex

Explore the impact of environmental conditions on marathon performance, and whether the impact differs across age and gender.

We first check the Pearson correlation among different weather condition variables. We depart WBGT into dry bulb, wet bulb, and black globe temperature. As shown in the plot, dry bulb, wet bulb, and black globe temperature are highly correlated with each other. We can use WBGT to represent the combination of the three types, containing information on temperature, humidity and solar radiation. Dew Point has high correlation with dry bulb and wet bulb temperature. The correlation between solar radiation and black globe temperature achieves 0.56. The correlation between solar radiation and relative humidity reaches -0.48. To explore the impact of environmental conditions on marathon performance, we first choose WBGT, AQI and wind speed, exploring their effects on marathon performance across age and gender, separately. We use the binary form of AQI and wind speed for better visualization. The threshold for AQI is 50, and 10 km/hour for wind speed.

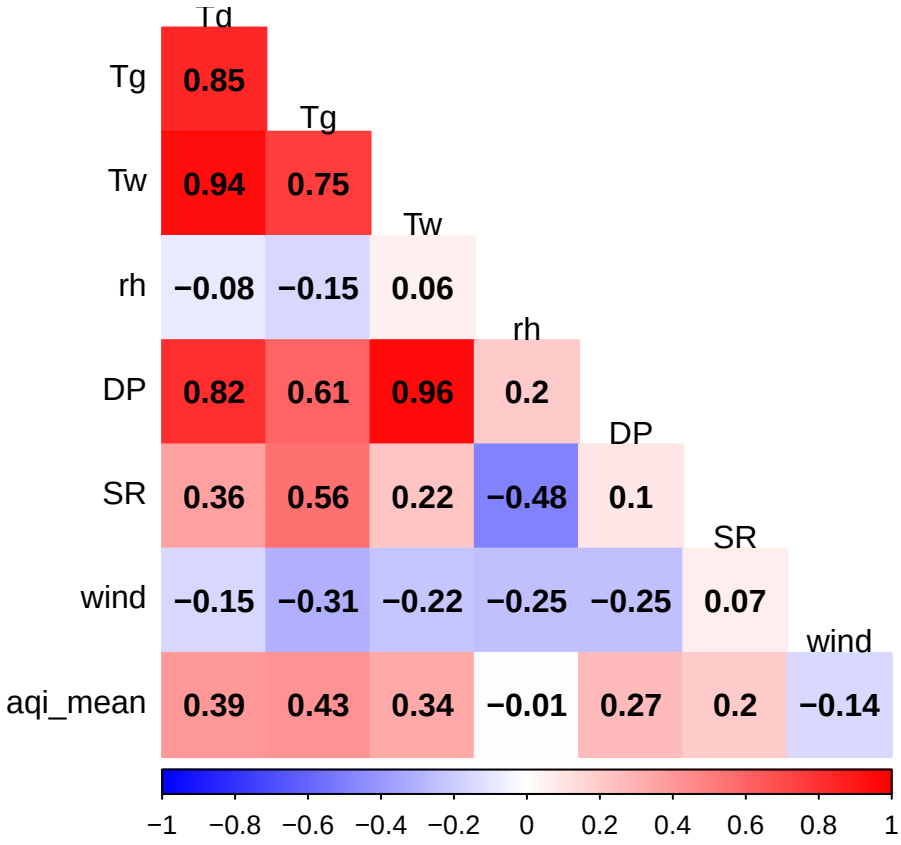

Figure 2: cor_matrix_plot

Wet Bulb Globe Temperature (WBGT) In the first set of graphs, we explore the effect of WBGT on marathon performance across age and gender. The results show that higher WBGT levels (represented by the red and yellow lines for "Red" and "Yellow" flag conditions) are associated with longer best finishing time, especially between 20 and 80 years of old, indicating that warmer environmental conditions negatively affect performance. This effect is more obvious at older ages, where performance deteriorates more quickly with increasing WBGT. One possible explanation is that there are only a few participants recorded above 80 years old under red WBGT flag weather condition. Thus, when WBGT>28°C, older people are more likely tend not to participate the race. For both men and women, WBGT seems to have a stronger impact as athletes age, particularly after age 60, where the gap between different WBGT levels widens.

Air Quality Index (AQI) The second set of graphs examines the impact of AQI on marathon performance, using AQI as a binary variable: "Good" (AQI < 50) and "Moderate" (AQI 50-100). The results suggest that AQI has a less noticeable effect on performance compared to WBGT. Both men and women show similar performance trends across age groups regardless of AQI levels, with only a slight difference between "Good" and "Moderate" AQI conditions. This suggests that air quality, within the observed range, does not drastically impact marathon performance. However, the lines slightly diverge at older ages, but the reasons might also be as described above.

Wind Speed The final set of graphs explores the effect of wind speed on performance, using a binary variable: wind speeds greater than 10 km/hour versus less than or equal to 10 km/hour. Similar to AQI, wind speed shows only a modest effect on performance. Both male and female participants experience slight performance differences based on wind conditions. We notice that for female, higher wind speeds leading to shorter best finishing times at older age groups (age > 60). As we do not the direction of the wind, one possible explanation is that a certain level of wind can make participants feel comfortable while running.

Overall, WBGT appears to be the most significant environmental factor affecting marathon performance, while AQI and wind speed have smaller effects.

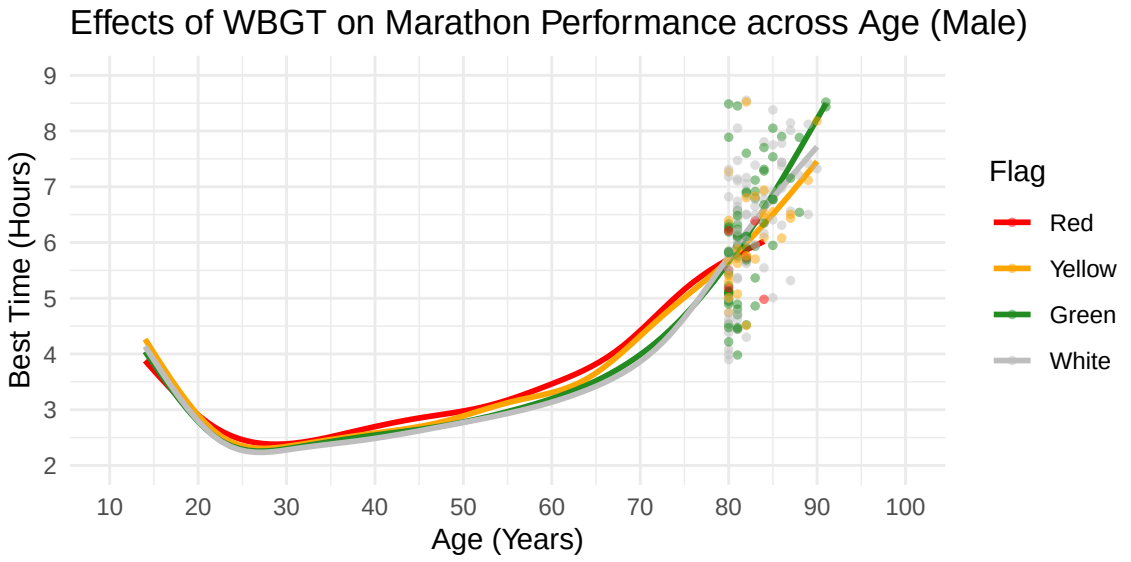

Figure 3: Effects of WBGT on Marathon Performance across Age (Male)

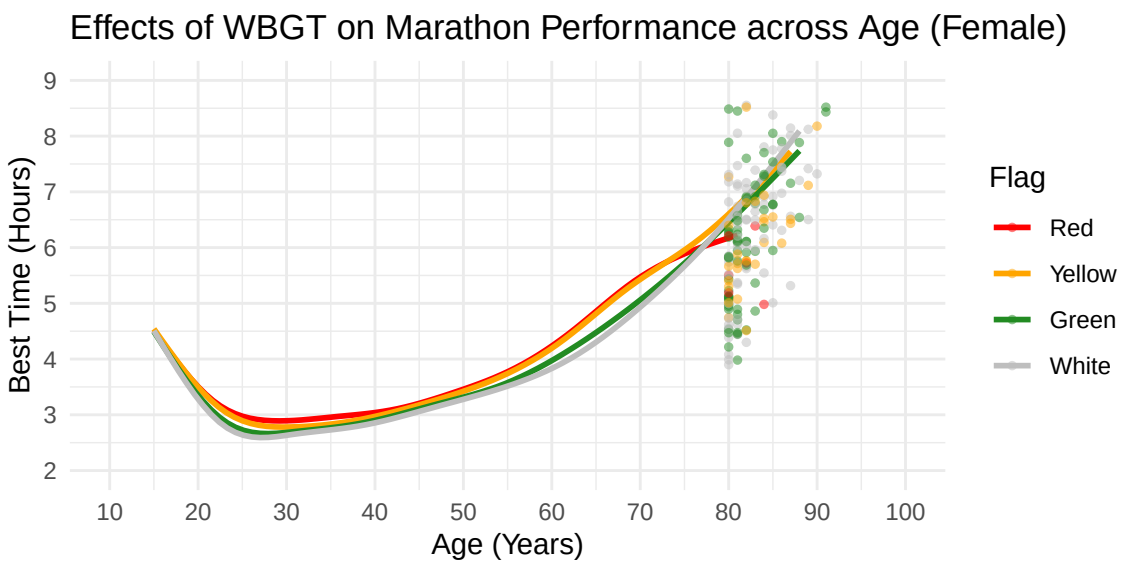

Figure 4: Effects of WBGT on Marathon Performance across Age (Female)

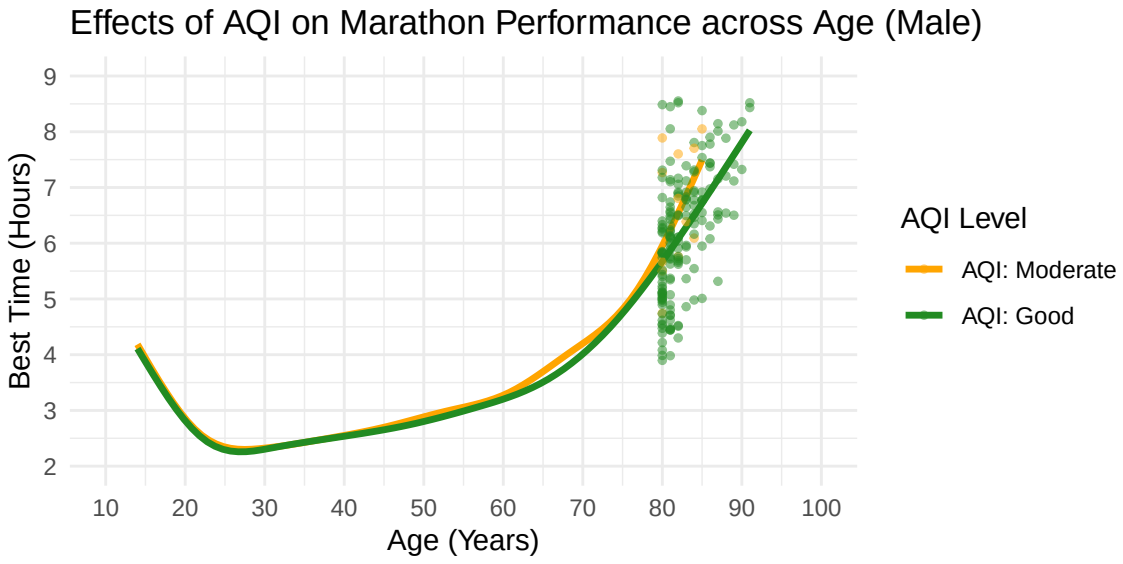

Figure 5: Effects of AQI on Marathon Performance across Age (Male)

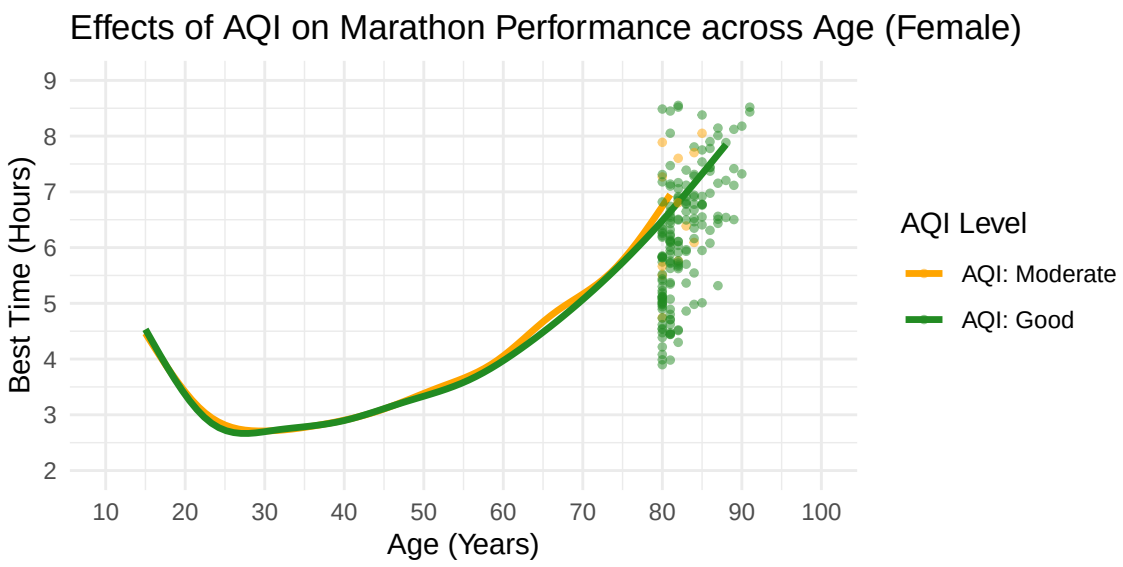

Figure 6: Effects of AQI on Marathon Performance across Age (Female)

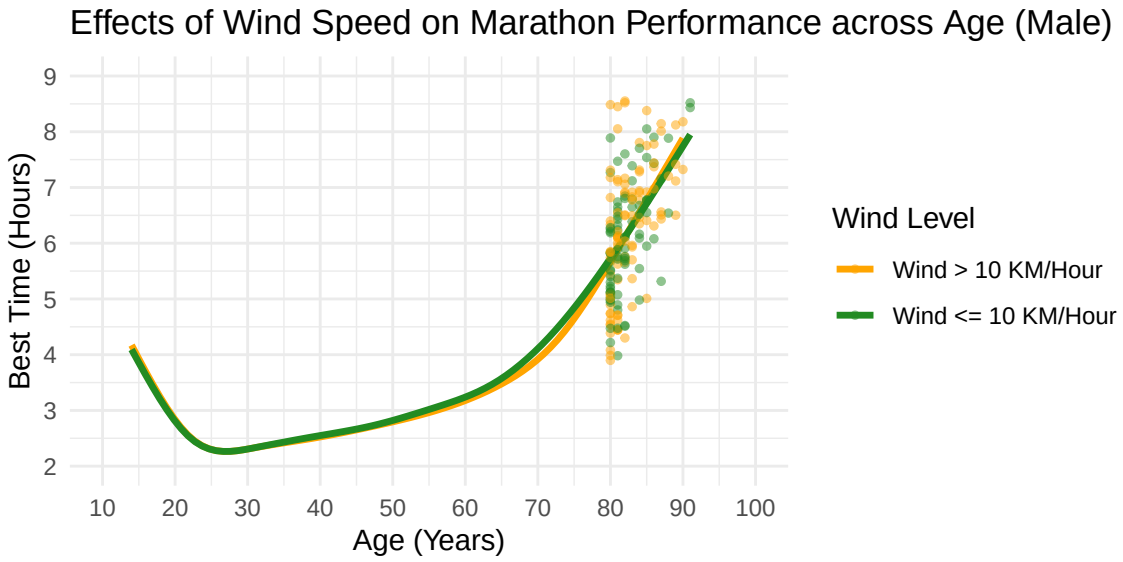

Figure 7: Effects of Wind Speed on Marathon Performance across Age (Male)

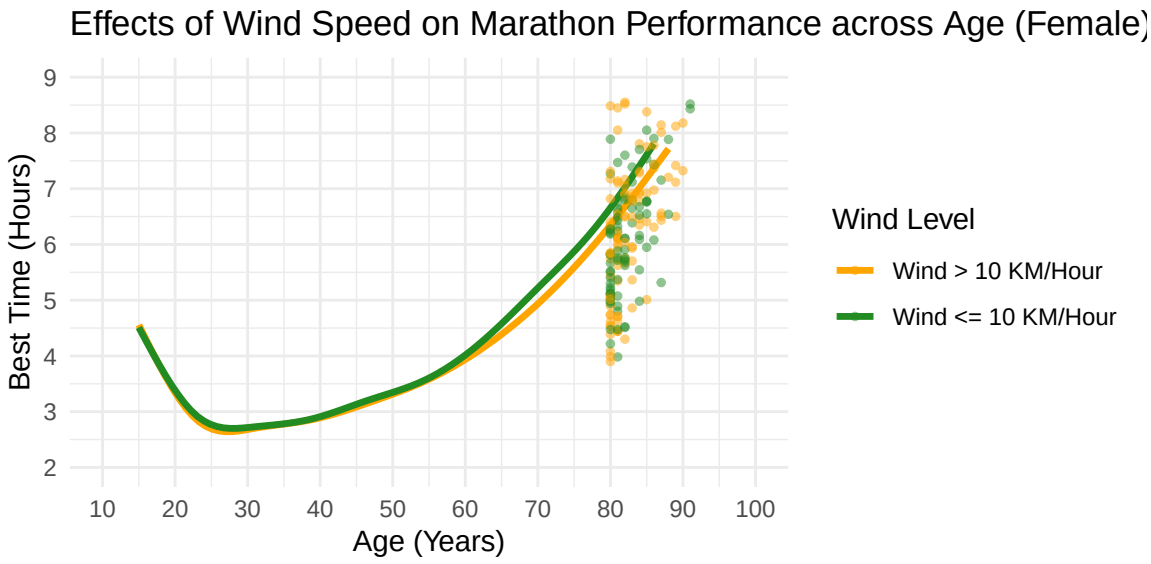

Figure 8: Effects of Wind Speed on Marathon Performance across Age (Female)

Identify the weather parameters (WBGT, Flag conditions, temperature, etc) that have the largest impact on marathon performance.

To identify the weather parameters that have the largest impact on marathon performance, we try to do some simple linear regression models for exploratory analysis. Our outcome of interest is Best finishing time (count in hours). For easy description, we call dry bulb temperature, wet bulb temperature, black globe temperature, relative humidity, wind speed, AQI these six varibales as selected weather condition variables.

Model 1 is constructed with predictors sex, age, selected weather condition variables and the interaction terms of sex with each selected weather condition variables, and the interactions terms of age with each selected weather condition variables. Results are shown in the table.

Model 2 is constructed all the same as model 1, except that it adds a square term of age

From the results table, we can learn that the coefficients for sex, age and aqi are significant in both models. The significance for the dry bulb temperature, wet bulb temperature, black globe temperature and relative humidity changes from model 1 to model 2. Thus, the form of the model has a large impact on our final results. It's hard to contain all the forms of variables and their different interactions in a single linear regression model.

Table 4: Linear Regression Model 1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.1990	0.1250	9.5896	0.0000
sex1	-0.5931	0.0865	-6.8559	0.0000
age	0.0530	0.0024	21.7819	0.0000
Td	-0.0465	0.0149	-3.1155	0.0018
Tw	0.0515	0.0130	3.9653	0.0001
Tg	0.0283	0.0056	5.0074	0.0000
rh	0.0029	0.0007	4.1387	0.0000
wind	0.0018	0.0058	0.3167	0.7515
aqi_mean	-0.0032	0.0015	-2.1081	0.0350
sex1:Td	-0.0028	0.0103	-0.2747	0.7835

	Estimate	Std. Error	t value	Pr(>|t|)
sex1:Tw	-0.0037	0.0090	-0.4175	0.6763
sex1:Tg	0.0026	0.0039	0.6590	0.5099
sex1:rh	0.0010	0.0005	2.0141	0.0440
sex1:wind	0.0041	0.0039	1.0342	0.3011
sex1:aqi_mean	0.0007	0.0011	0.6842	0.4939
age:Td	0.0013	0.0003	4.5948	0.0000
age:Tw	-0.0010	0.0003	-3.9129	0.0001
age:Tg	-0.0007	0.0001	-6.4055	0.0000
age:rh	-0.0001	0.0000	-6.3115	0.0000
age:wind	-0.0002	0.0001	-1.7384	0.0822
age:aqi_mean	0.0000	0.0000	0.6484	0.5168

Table 5: Linear Regression Model 2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.9783	0.0857	58.1177	0.0000
sex1	-0.6993	0.0555	-12.6000	0.0000
age	-0.1225	0.0021	-58.5307	0.0000
Td	-0.0064	0.0096	-0.6700	0.5029
Tw	0.0137	0.0083	1.6493	0.0991
Tg	0.0064	0.0036	1.7734	0.0762
rh	0.0008	0.0004	1.6973	0.0897
wind	0.0068	0.0037	1.8294	0.0674
aqi_mean	-0.0020	0.0010	-2.0872	0.0369
I(age^2)	0.0018	0.0000	125.7167	0.0000
sex1:Td	-0.0005	0.0066	-0.0784	0.9375
sex1:Tw	-0.0058	0.0058	-1.0045	0.3152
sex1:Tg	0.0017	0.0025	0.6778	0.4979
sex1:rh	0.0011	0.0003	3.5698	0.0004
sex1:wind	0.0032	0.0025	1.2643	0.2062
sex1:aqi_mean	0.0012	0.0007	1.7045	0.0883
age:Td	0.0002	0.0002	1.1256	0.2604
age:Tw	0.0000	0.0002	0.1140	0.9092
age:Tg	-0.0001	0.0001	-2.0120	0.0442
age:rh	0.0000	0.0000	-4.8266	0.0000
age:wind	-0.0003	0.0001	-3.6313	0.0003
age:aqi_mean	0.0000	0.0000	0.1112	0.9114

We decide to fit a random forest model instead. This is because random forest is able to consider all the forms of variables and their interactions. We can simply input all the predictors into the model, since random forest itself is able to choose among the predictors.

Our outcome of interest is still Best finishing time (count in hours). We input dry bulb temperature, wet bulb temperature, black globe temperature, relative humidity, wind speed, AQI, solar radiation, dew point, WBGT, and WBGT flag into the model.

As shown in the variable importance plot, the left panel - %IncMSE (Percentage Increase in Mean Squared Error) - represents how much the model's mean squared error (MSE) increases when a particular variable is randomly permuted while others remain unchanged. The right panel - IncNodePurity (Increase in Node Purity) - shows the increase in node purity, which reflects how much a variable improves the homogeneity of the nodes and leaves in the trees of the Random Forest.

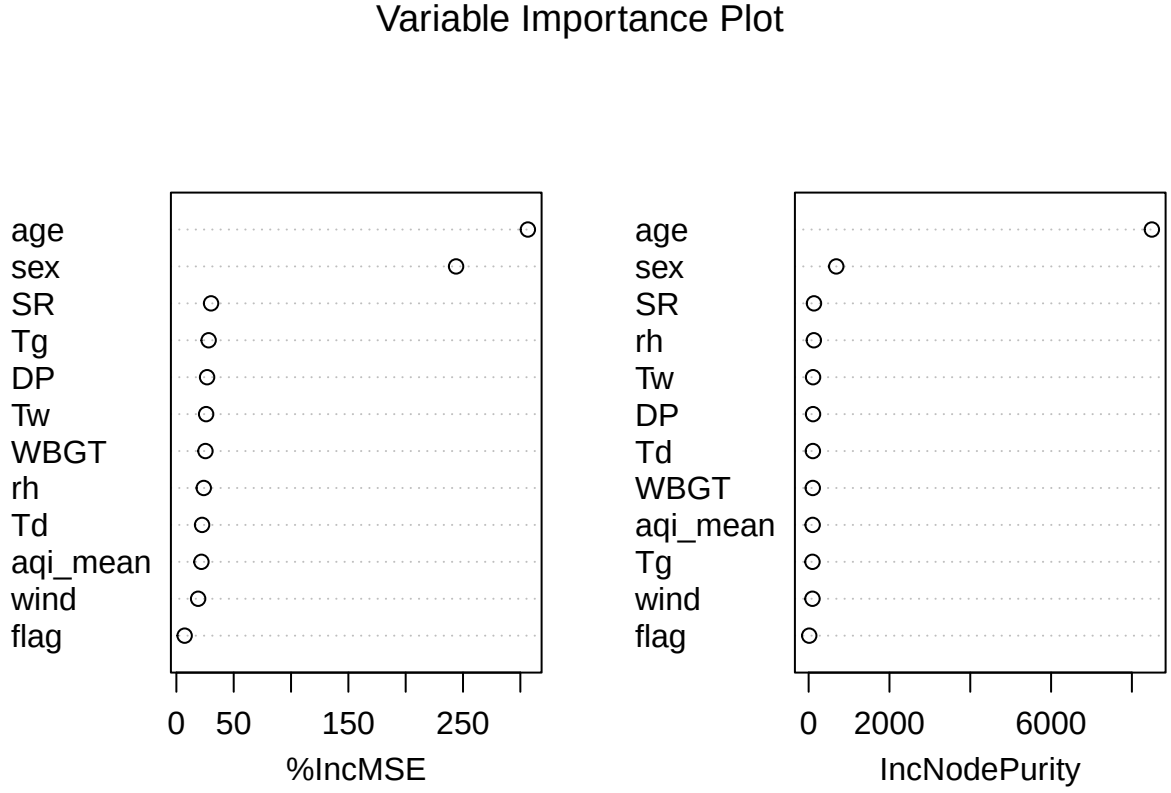

Figure 9: Variable Importance Plot

We can learn that age and sex are the most important predictors in explaining the marathon performance, while other weather condition variables have small but similar contribution in explaining the marathon performance.

Discussion

In this study, we explored how environmental conditions impact marathon performance across age and sex. Our results indicate that age and sex are the most significant predictors of performance, with men generally performing faster than women across all age groups, and marathon times increasing with age, particularly after the age of 60.

Among the environmental factors, WBGT (Wet Bulb Globe Temperature) had the most pronounced effect when not adjusting for other factors, as higher temperatures led to longer finishing times. However, AQI (Air Quality Index) and wind speed had only modest effects on performance, with minor differences observed across the binary categories of these variables. The effect of these weather conditions is more obvious at older ages, but this may be due to the reason that there are only a few old participants participate under "bad" weather conditions. The small sample size for old age group may lead any conclusions not reliable. One interesting finding is that for female, higher wind speeds leading to shorter best finishing times at older age groups (age > 60).

When identifying the weather condition factor that has the most impact on marathon performance, the form of variables and the interactions among them is very complicated. Given all the weather condition factors we have, a random forest model tells us that the contribution in explaining the performance are similar among all the weather condition factors.

One limitation of our study is the lack of detailed information on wind direction, which could provide further insights into the impact of wind on performance. Another limitation is the absence of data on participants who failed to complete the marathon. Our outcome of interest is the best finishing time in each age-gender group, which may be robust against changes in weather conditions, as these runners are already the best in their group and may not be easily affected by environmental factors.

Conclusion

Overall, we conclude that both men and women reach their best marathon performance around their mid-20s. After reaching this peak, performance gradually declines with increasing age. While age and sex are the dominant factors influencing marathon performance, WBGT has a more significant effect on performance compared to other weather conditions, particularly in older participants, when not adjusting for other factors. A similar trend is observed in both males and females. However, when adjusting for other factors, a random forest model shows that the effects of weather conditions, such as WBGT, wind speed, and AQI, though present, are less substantial.

Appendix

```
library(tidyverse)
library(lubridate)
library(corrplot)
library(randomForest)
library(knitr)
library(kableExtra)
library(gtsummary)

aqi = read.csv(
```

```
"D:/Brown/2024fall/Course/PHP2550 Practical Data Analysis/Project/Project1/aqi_values.csv"
)
course_record = read.csv(
"D:/Brown/2024fall/Course/PHP2550 Practical Data Analysis/Project/Project1/course_record.csv"
)
marathon_dates = read.csv(
"D:/Brown/2024fall/Course/PHP2550 Practical Data Analysis/Project/Project1/marathon_dates.csv"
)
dat.orig = read.csv(
"D:/Brown/2024fall/Course/PHP2550 Practical Data Analysis/Project/Project1/project1.csv"
)

dat = dat.orig %>%
  rename(
    race = Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D.,
    year = Year,
    sex = Sex..0.F..1.M.,
    flag = Flag,
    age = Age..yr.,
    CR = X.CR,
    Td = Td..C,
    Tw = Tw..C,
    rh = X.rh,
    Tg = Tg..C,
    SR = SR.W.m2,
    wind = Wind
  ) %>%
  mutate(flag = factor(flag,
                       levels = c("", "Red", "Yellow", "Green", "White")),
         sex = as.factor(sex))

dim(dat)
# Calculating the best time in hours
dat_cr = course_record %>%
  mutate(race = case_when(
    Race == "B" ~ 0,
    Race == "C" ~ 1,
    Race == "NY" ~ 2,
    Race == "TC" ~ 3,
    Race == "D" ~ 4
  )) %>%
  mutate(sex = case_when(Gender == "M" ~ 1, Gender == "F" ~ 0)) %>%
  mutate(sex = as.factor(sex)) %>%
  mutate(CRtime = period_to_seconds(hms(CR))) %>%
  mutate(CRtime = as.numeric(CRtime)) %>%
  rename(year = Year) %>%
  select(race, year, sex, CRtime)

anti_join(dat_cr,
          dat %>% distinct(race, year, sex),
          by = c("race", "year", "sex"))

dat = left_join(dat, dat_cr, by = c("race", "year", "sex"))
```

```r
dat = dat %>% mutate(besttime = CRtime * (1 + CR / 100) / 3600)
# Process AQI (Air Quality Index)
aqi.process = aqi %>%
  mutate(year = year(ymd(date_local))) %>%
  mutate(year = as.numeric(year)) %>%
  mutate(
    race = case_when(
      marathon == "Boston" ~ 0,
      marathon == "Chicago" ~ 1,
      marathon == "NYC" ~ 2,
      marathon == "Twin Cities" ~ 3,
      marathon == "Grandmas" ~ 4
    )
  )

aqi.process = aqi.process %>%
  group_by(race, year) %>%
  summarise(aqi_mean = mean(aqi, na.rm = T))

dat = left_join(dat, aqi.process, by = c("race", "year"))

range(dat$aqi_mean)
sum(is.na(dat$aqi_mean))
hist(dat$aqi_mean)
range(dat$wind)

dat = dat %>%
  mutate(aqi_good = ifelse(aqi_mean <= 50, 1, 0)) %>%
  # 1 means good quality, 0 means moderate quality
  mutate(wind_lt_10 = ifelse(wind <= 10, 1, 0))
# Learn the data
table(dat$race)
#0 = Boston Marathon
#1 = Chicago Marathon
#2 = New York City Marathon
#3 = Twin Cities Marathon (Minneapolis, MN)
#4= Grandma's Marathon (Duluth, MN)
table(dat$year)
table(dat$sex) #0=female, 1=male
table(dat$flag)
range(dat$age); hist(dat$age,xlim=c(0,100), breaks=20)
range(dat$CR); hist(dat$CR,xlim = c(-5,420), breaks=85)
head(sort(dat$CR, decreasing=T))
range(dat$Td, na.rm=T); hist(dat$Td)
range(dat$Tw, na.rm=T); hist(dat$Tw)
range(dat$rh, na.rm=T); hist(dat$rh)
range(dat$Tg, na.rm=T); hist(dat$Tg)
range(dat$SR, na.rm=T); hist(dat$SR)
range(dat$DP, na.rm=T); hist(dat$DP)
range(dat$wind, na.rm=T); hist(dat$wind)
range(dat$WBGT, na.rm=T); hist(dat$Td)

dat %>% distinct(race, year)
```

```r
table_race_year = dat %>% group_by(race) %>%
  summarise(number_of_year = n_distinct(year),
            min_year = min(year),
            max_year = max(year))
table_race_year
dat %>% filter(race==0) %>% distinct(year) # miss year 1999
dat %>% filter(race==1) %>% distinct(year)
dat %>% filter(race==2) %>% distinct(year) # miss year 2012
dat %>% filter(race==3) %>% distinct(year)
dat %>% filter(race==4) %>% distinct(year)
# Plot the table for Marathon Race Frequency and Year Range
table_race_year = table_race_year %>%
  mutate(
    race = case_when(
      race == 0 ~ "Boston",
      race == 1 ~ "Chicago",
      race == 2 ~ "NYC",
      race == 3 ~ "Twin Cities",
      race == 4 ~ "Grandma"
    )
  )
colnames(table_race_year) = c("Race", "Number of Times", "Min Year", "Max Year")
table_race_year %>% kbl(caption = "Marathon Race Frequency and Year Range Overview") %>%
  kable_styling(
    bootstrap_options = c("striped", "hover", "condensed"),
    full_width = F,
    position = "center"
  )
# Missing data checking
apply(dat, 2, function(x)
  sum(is.na(x)))
table(dat$flag)

apply(dat[, c("Td", "Tw", "rh", "Tg", "SR", "DP", "wind", "WBGT")], 2, function(x)
  range(x, na.rm = T))
apply(dat[, c("Td", "Tw", "rh", "Tg", "SR", "DP", "wind", "WBGT")], 2, function(x)
  hist(x))

flag.missing = dat %>% filter(flag == "")
apply(flag.missing[, c("Td", "Tw", "rh", "Tg", "SR", "DP", "wind", "WBGT")],
      2, function(x)
  sum(is.na(x)))
flag.missing %>%
  select(race, year) %>%
  distinct()
flag.missing %>% group_by(race, year, sex) %>% summarise(mean_CR = mean(CR))
# Drop missing values and only remain complete cases
dat = dat %>% filter(!flag=="")
# Plot the table for the Distribution of age and sex across races
dat = dat %>%
  mutate(
    age_bin = case_when(
      age < 20 ~ 1,
```

```
      age >= 20 & age < 30 ~ 2,
      age >= 30 & age < 40 ~ 3,
      age >= 40 & age < 50 ~ 4,
      age >= 50 & age < 60 ~ 5,
      age >= 60 & age < 70 ~ 6,
      age >= 70 & age < 80 ~ 7,
      age >= 80 ~ 8
    )
  ) %>%
  mutate(age_bin = factor(age_bin))

dat %>% select(race, sex, age_bin) %>%
  mutate(
    race = case_when(
      race == 0 ~ "Boston",
      race == 1 ~ "Chicago",
      race == 2 ~ "NYC",
      race == 3 ~ "Twin Cities",
      race == 4 ~ "Grandma"
    ),
    sex = case_when(sex == 0 ~ "Female", sex == 1 ~ "Male"),
    age_bin = case_when(
      age_bin == 1 ~ "0-19",
      age_bin == 2 ~ "20-29",
      age_bin == 3 ~ "30-39",
      age_bin == 4 ~ "40-49",
      age_bin == 5 ~ "50-59",
      age_bin == 6 ~ "60-69",
      age_bin == 7 ~ "70-79",
      age_bin == 8 ~ "80-100"
    )
  ) %>%
  tbl_summary(.,
              by = race,
              label = list(sex ~ "**Gender**", age_bin ~ "**Age Group**")) %>%
  add_p() %>%
  modify_header(label = "**Variables**") %>%
  modify_caption("**Distribution of Sex and Age across Races**") %>%
  as_kable()
age_sex_plot =
  ggplot(dat, aes(x = age, y = besttime, color = factor(sex))) +
  geom_smooth(size = 2, se = FALSE) +
  geom_point(alpha = 0.1,
             size = 1,
             position = position_jitter(width = 0.2)) +
  labs(
    title = "Effects of Increasing Age on Marathon Performance across Sex",
    x = "Age (Years)",
    y = "Best Time (Hours)",
    color = "Sex"
  ) +
  scale_color_discrete(labels = c("Female", "Male")) +
  theme_minimal()
```

```r
age_sex_plot
ggsave("../Plots_Project1/age_sex_plot.png", plot = age_sex_plot, width = 10, height = 8)
# Plot Correlation Matrix among Weather Condition Variables
cor_matrix <- cor(dat[, c("Td", "Tg", "Tw", "rh", "DP", "SR", "wind", "aqi_mean")])

corrplot(
  cor_matrix,
  method = "color",
  col = colorRampPalette(c("blue", "white", "red"))(200),
  type = "lower",
  addCoef.col = "black",
  tl.col = "black",
  tl.srt = 0,
  diag = FALSE
)
# Learn the number and proportion of each WBGT flag in all the races
dat %>%
  distinct(race, year, flag)
#White = WBGT <10C
#Green = WBGT 10-18C
#Yellow = WBGT >18-23C
#Red = WBGT >23-28C
#Black = WBGT >28C
dat %>%
  distinct(race, year, flag) %>%
  mutate(flag = factor(flag, levels = c("Red", "Yellow", "Green", "White"))) %>%
  group_by(flag) %>%
  summarise(count = n()) %>%
  mutate(prop = count / sum(count))
# Effects of WBGT on marathon performance across age and gender
dat_male = dat %>% filter(sex == 1)
dat_female = dat %>% filter(sex == 0)

ggplot(dat_male, aes(x = age, y = besttime, color = factor(flag))) +
  geom_smooth(se = FALSE) +
  geom_point(
    data = dat_male %>% filter(age >= 80),
    aes(x = age, y = besttime, color = factor(flag)),
    alpha = 0.5,
    size = 1
  ) +
  labs(
    title = "Effects of WBGT on Marathon Performance across Age (Male)",
    x = "Age (Years)",
    y = "Best Time (Hours)",
    color = "Flag"
  ) +
  scale_color_manual(values = c(
    "Red" = "red",
    "Yellow" = "orange",
    "Green" = "forestgreen",
    "White" = "grey"
  )) +
```

```
  scale_x_continuous(limits = c(10, 100), breaks = seq(10, 100, by = 10)) +
  scale_y_continuous(limits = c(2, 9), breaks = seq(2, 9, by = 1)) +
  theme_minimal()
ggplot(dat_female, aes(x = age, y = besttime, color = factor(flag))) +
  geom_smooth(se = FALSE) +
  geom_point(
    data = dat_male %>% filter(age >= 80),
    aes(x = age, y = besttime, color = factor(flag)),
    alpha = 0.5,
    size = 1
  ) +
  labs(
    title = "Effects of WBGT on Marathon Performance across Age (Female)",
    x = "Age",
    y = "Performance (CR)",
    color = "Flag"
  ) +
  scale_color_manual(values = c(
    "Red" = "red",
    "Yellow" = "orange",
    "Green" = "forestgreen",
    "White" = "grey"
  )) +
  scale_x_continuous(
    name = "Age (Years)",
    limits = c(10, 100),
    breaks = seq(10, 100, by = 10)
  ) +
  scale_y_continuous(name = "Best Time (Hours)",
                     limits = c(2, 9),
                     breaks = seq(2, 9, by = 1)) +
  theme_minimal()
# Effects of AQI on marathon performance across age and gender
ggplot(dat_male, aes(
  x = age,
  y = besttime,
  color = factor(aqi_good)
)) +
  geom_smooth(se = FALSE, size = 1.2) +
  geom_point(
    data = dat_male %>% filter(age >= 80),
    aes(
      x = age,
      y = besttime,
      color = factor(aqi_good)
    ),
    alpha = 0.5,
    size = 1
  ) +
  labs(
    title = "Effects of AQI on Marathon Performance across Age (Male)",
    x = "Age (Years)",
    y = "Best Time (Hours)",
```

```
    color = "AQI Level"
  ) +
  scale_color_manual(
    values = c("1" = "forestgreen", "0" = "orange"),
    labels = c("1" = "AQI: Good", "0" = "AQI: Moderate")
  ) +
  scale_x_continuous(limits = c(10, 100), breaks = seq(10, 100, by = 10)) +
  scale_y_continuous(limits = c(2, 9), breaks = seq(2, 9, by = 1)) +
  theme_minimal()
#Good (0-50): Air quality is considered satisfactory
#Moderate (51-100): Air quality is acceptable

ggplot(dat_female, aes(
  x = age,
  y = besttime,
  color = factor(aqi_good)
)) +
  geom_smooth(se = FALSE, size = 1.2) +
  geom_point(
    data = dat_male %>% filter(age >= 80),
    aes(
      x = age,
      y = besttime,
      color = factor(aqi_good)
    ),
    alpha = 0.5,
    size = 1
  ) +
  labs(
    title = "Effects of AQI on Marathon Performance across Age (Female)",
    x = "Age (Years)",
    y = "Best Time (Hours)",
    color = "AQI Level"
  ) +
  scale_color_manual(
    values = c("1" = "forestgreen", "0" = "orange"),
    labels = c("1" = "AQI: Good", "0" = "AQI: Moderate")
  ) +
  scale_x_continuous(limits = c(10, 100), breaks = seq(10, 100, by = 10)) +
  scale_y_continuous(limits = c(2, 9), breaks = seq(2, 9, by = 1)) +
  theme_minimal()

# Learn the distribution of wind speed
dat %>%
  distinct(race, year, wind) %>%
  summarise(
    count_wind_large_than_20 = sum(wind > 20),
    count_wind_10_20 = sum(wind > 10 & wind <= 20),
    count_wind_0_10 = sum(wind <= 10)
  )
# Effects of wind speed on marathon performance across age and gender
ggplot(dat_male, aes(
  x = age,
```

```
  y = besttime,
  color = factor(wind_lt_10)
)) +
  geom_smooth(se = FALSE, size = 1.2) +
  geom_point(
    data = dat_male %>% filter(age >= 80),
    aes(
      x = age,
      y = besttime,
      color = factor(wind_lt_10)
    ),
    alpha = 0.5,
    size = 1
  ) +
  labs(
    title = "Effects of Wind Speed on Marathon Performance across Age (Male)",
    x = "Age (Years)",
    y = "Best Time (Hours)",
    color = "Wind Level"
  ) +
  scale_color_manual(
    values = c("1" = "forestgreen", "0" = "orange"),
    labels = c("1" = "Wind <= 10 KM/Hour", "0" = "Wind > 10 KM/Hour")
  ) +
  scale_x_continuous(limits = c(10, 100), breaks = seq(10, 100, by = 10)) +
  scale_y_continuous(limits = c(2, 9), breaks = seq(2, 9, by = 1)) +
  theme_minimal()


ggplot(dat_female, aes(
  x = age,
  y = besttime,
  color = factor(wind_lt_10)
)) +
  geom_smooth(se = FALSE, size = 1.2) +
  geom_point(
    data = dat_male %>% filter(age >= 80),
    aes(
      x = age,
      y = besttime,
      color = factor(wind_lt_10)
    ),
    alpha = 0.5,
    size = 1
  ) +
  labs(
    title = "Effects of Wind Speed on Marathon Performance across Age (Female)",
    x = "Age (Years)",
    y = "Best Time (Hours)",
    color = "Wind Level"
  ) +
  scale_color_manual(
    values = c("1" = "forestgreen", "0" = "orange"),
```

```
    labels = c("1" = "Wind <= 10 KM/Hour", "0" = "Wind > 10 KM/Hour")
  ) +
  scale_x_continuous(limits = c(10, 100), breaks = seq(10, 100, by = 10)) +
  scale_y_continuous(limits = c(2, 9), breaks = seq(2, 9, by = 1)) +
  theme_minimal()
m1 = lm(
  besttime ~ sex + age + Td + Tw + Tg + rh + wind + aqi_mean +
    Td:sex + Tw:sex + Tg:sex + rh:sex + wind:sex + aqi_mean:sex +
    Td:age + Tw:age + Tg:age + rh:age + wind:age + aqi_mean:age,
  data = dat
)
kable(round(summary(m1)$coefficients, 4), caption = "Linear Regression Model 1")

m2 = lm(
  besttime ~ sex + age + Td + Tw + Tg + rh + wind + aqi_mean + I(age ^ 2) +
    Td:sex + Tw:sex + Tg:sex + rh:sex + wind:sex + aqi_mean:sex +
    Td:age + Tw:age + Tg:age + rh:age + wind:age + aqi_mean:age,
  data = dat
)
kable(round(summary(m2)$coefficients, 4), caption = "Linear Regression Model 2")
summary(m1)
summary(m2)
# random forest model
start = Sys.time()
set.seed(1)
rf_m = randomForest(
  x = dat[, c("sex",
              "age",
              "Td",
              "Tw",
              "Tg",
              "rh",
              "wind",
              "aqi_mean",
              "SR",
              "DP",
              "WBGT",
              "flag")],
  y = dat[, "besttime"],
  keep.forest = TRUE,
  importance = TRUE
)
end = Sys.time()
end - start
print(rf_m)
rf_m$importance
varImpPlot(rf_m, main="Variable Importance Plot")
```