Project 1: Exploratory data analysis of environmental conditions and marathon performance

Due: October 6 2024

Miaoyan Chen

Introduction

The purpose of Dr. Brett Romano Ely and Dr. Matthew Ely's research is to examine the impact of weather on marathon performance across lifespan in both men and women. Previous preliminary result shows that the performance in marathons is degraded with increasing environmental temperature, and the impact is more significant for longer-distant events (ELY et al. 2007). In addition, older runners are more sensitive to heat stress than younger runners (Kenney and Munce 2003). There are also well-documented difference in endurance performance between genders (Besson et al. 2022). In this exploratory data analysis, we will investigate the impact of environmental conditions (e.g. temperature, humidity, solar radiation, and wind) on marathon performance between men and women. Dr. Ely hypothesized that slowing would be more pronounced with high WBGT in older individuals compared to younger individuals and similar patterns would be observed in men and women.

I will be using the marathon data set to conduct a series of exploratory data analysis to examine the effects of increasing age on marathon performance in men and women. I will also explore the impact of environmental conditions on marathon performance, and whether the impact differs across age and gender. Lastly, I will identify the weather parameters that have the largest impact on marathon performance. I will begin by conducting a data quality check and identify any missing data patterns before proceeding with the exploratory data analysis.

Data quality, cleaning and missingness

The marathon dataset contains 14 parameters. `race`, `year`,`sex`, and `flag` are recorded as categorical variable. While age and other weather parameters such as `percent off course record`, `dry bulb temperature`, `wet bulb temperature`, `black globe temperature`, `solar radiation`, `dew point`, `wind speed`, and `Wet Bulb Globe Temperature` are continuous variables. The data set includes top single age performances from 5 major marathons (Boston, Chicago, New York City (NYC), Twin Cities, and Duluth) across years 1993 to 2016 for men and women. Each marathon race has detailed weather and environmental conditions recorded in the data set.

Overall, the data is well structured and data code book explain what each variable in the data set represents. To begin the data cleaning process, I factored the aforementioned categorical variables and ensure each of the weather parameters are numerically coded in the data set. I renamed the variables to make tables and figures more understandable. I then merged the marathon data set with the course record data set to calculate the actual course record (finish time) for each runner. The actual course record (finish time) is calculated by adding the percentage course record on top of the best course record for each runner. I also examined the distribution of the continuous variables and realize none of the environmental conditions are approximately normal distributed. Therefore, I would use the median instead of mean to summarize the percent off course record and runner's finish time in my analysis.

The only missingness in this data set is the missing observations in 2011 and 2012 marathon races. For example, the data set does not have any observations for Chicago, NYC, Twin Cities, and Grandma's (Duluth) marathon in 2011. The missingness in the data set is completely random (MCAR) and is likely due to the fact that the marathon was not held at this time in these four cities, or there's no weather data available. I will not address the missingness in the data set since the missingness is not due to data entry errors.

Characteristics of marathon runners and locations

Tables 1 and 2 summarizes runner's characteristics and environmental conditions for marathon locations. Table 1 is summarized by gender, and it includes observations across all ages, years, and marathon locations. The table shows that the actual course record is different between men and women. The median score for female runners is 5 minutes higher than male runners. The table also shows that female and male distributions are similar across all marathon locations, and the median age for male runners is a little higher than female runners.

Table 1: Characteristics of marathon runners

	Gender	
Characteristic	**Female** N = 5,452	**Male** N = 6,112
Actual course record (in minutes)	197 (171, 234)	192 (167, 232)
Location		
Boston	984 (18%)	1,104 (18%)
Chicago	1,210 (22%)	1,343 (22%)
Grandma's	934 (17%)	1,066 (17%)
New York City	1,402 (26%)	1,528 (25%)
Twin Cities	922 (17%)	1,071 (18%)
Age	45 (30, 59)	48 (32, 64)

[1] Median (Q1, Q3); n (%)

Table 2 summarizes the environmental conditions for each marathon location across all years. The table shows that there's variation in all the weather parameters among marathon locations. In general, Grandma's marathon has the highest median temperature in dry bulb, wet bulb, WBGT, black globe, and dew point. It also has the highest solar radiation. Therefore, it could mean that Grandma's marathon has the highest heat stress among all marathon locations. Whereas, Boston marathon has the lowest median temperature in dry bulb and wet bulb. It also has the lowest solar radiation. Boston and New York City has relatively similar weather conditions as they are both located in the east coast. The table also shows that `Percent relative humidity` is very skewed with many observation less than or equal to 1, which could be due to data entry errors by not converting the values to percentage. I converted the these values less than 1 into percentage for further analysis.

Table 2: Environmental conditions for marathon locations

Measurements	Marathons					p-value
	Boston N = 18	Chicago N = 21	Grandma's N = 17	NYC N= 23	Twin Cities N = 17	
Dry bulb	9 (8, 14)	14 (7, 15)	19 (16, 22)	12 (7, 15)	12 (9, 16)	<0.001
Wet bulb	7 (5, 8)	9 (3, 13)	14 (14, 16)	7 (3, 12)	9 (7, 13)	<0.001
% relative humidity	62 (46, 73)	60 (53, 68)	64 (56, 84)	53 (44, 61)	63 (56, 75)	0.12
Wind speed	12 (8, 16)	8 (5, 10)	9 (8, 11)	11 (9, 14)	9 (7, 10)	0.010
WBGT	10 (9, 13)	13 (7, 16)	18 (16, 21)	10 (7, 14)	13 (9, 16)	<0.001
Black globe	23 (19, 28)	26 (21, 29)	34 (28, 38)	20 (18, 25)	26 (20, 30)	0.003
Solar radiation	721 (574, 800)	470 (437, 518)	736 (571, 838)	393 (309, 546)	488 (355, 541)	<0.001
Dew point	3 (0, 6)	6 (-2, 10)	12 (11, 14)	2 (-4, 9)	6 (3, 10)	<0.001

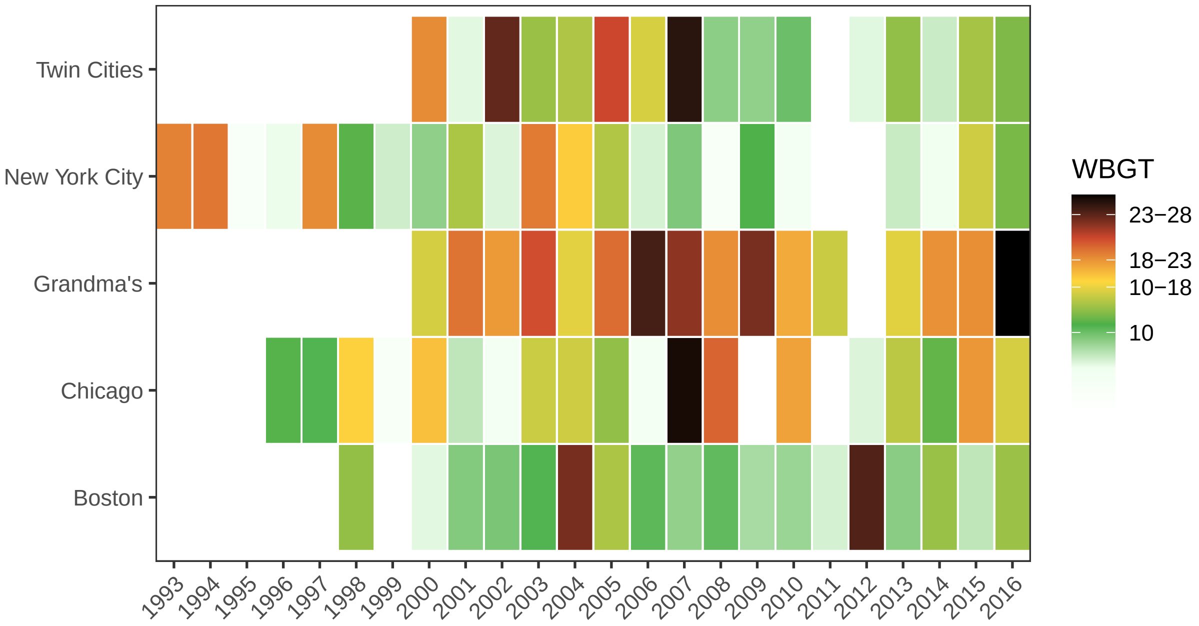
[1] Median (Q1, Q3)

[2] Kruskal-Wallis rank sum test

Exploratory data analysis

The colors of this heatmap is utilizes the levels of `Flag`, with warmer colors indicating warmer temperature. The figure shows that Wet Bulb Globe Temperature (WBGT) is highest in Grandma's marathon, followed by Twin Cities, Chicago, New York City, and Boston. Marathon locations located in the east coast, such as NYC and Boston shows lower WBGT compared to marathon locations in the midwest. This suggests that weather and environmental conditions are different across marathon races.

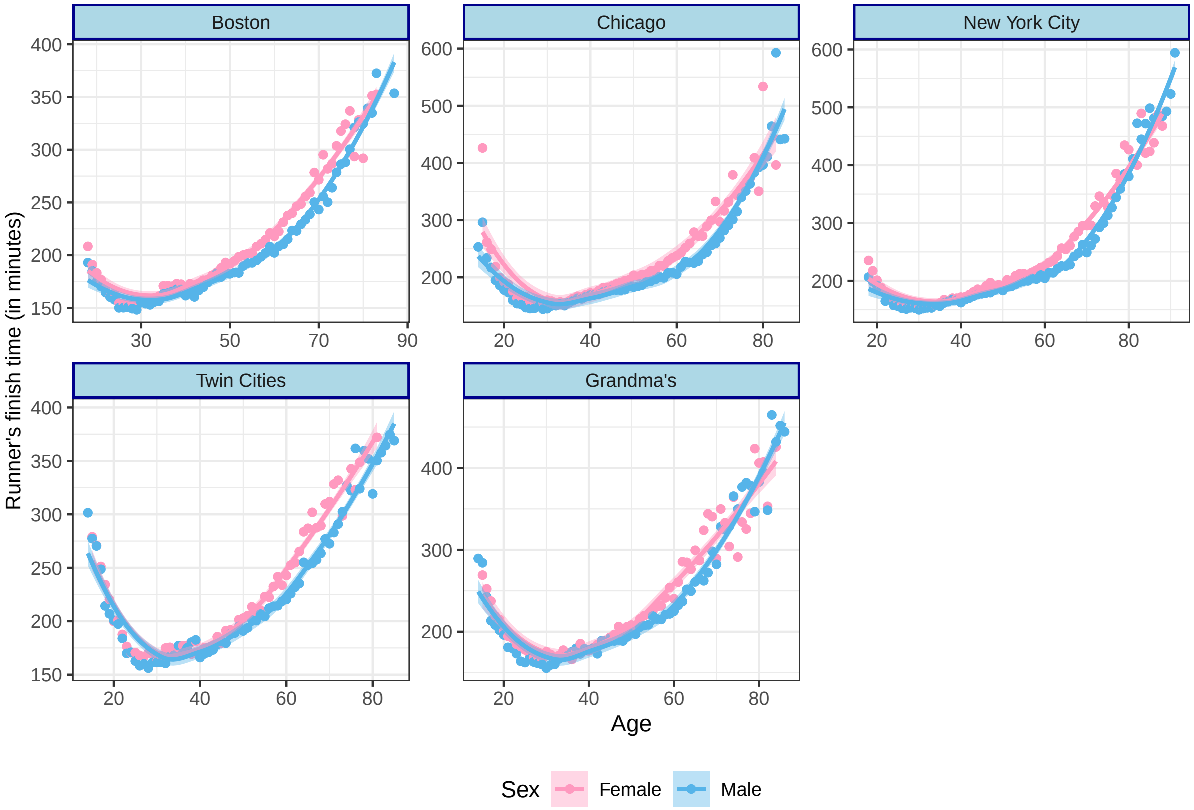
Figure 1: Heatmap of WBGT across marathons and years



The `marathon` data set is merged with the `course record` data set to calculate the actual marathon finish time for each runner. Each runner's actual marathon course record is calculated by adding on the percentage off course record (% CR) on top of the best course record for each race. Course record is the fastest time that a runner has completed a marathon, and therefore the lower the course record, the better the performance. The following scatter plot shows the association between the actual course record (in minutes) and age for male and female runner at each marathon location. The association between course record and age is quadratic for all marathons, which indicates that age is a significant factor that effects the outcome of the marathon. Between ages 14 -30, teenage and young adult runners runners are showing an increase in performance as they age. However, when runners reach around age 30, this is where the peak of the performance is observed. After age 30, the performance starts to decrease as runners age. The fastest runner for each is around age 30, and the performance decreases with increasing age. The corresponding table shows the fastest runner for each marathon in the past decade.

Aim 1: Examine effects of increasing age on marathon performance in men and women

Figure 2: Association of course record with age and gender



The 5 scatter plots (Figure 2) show that performance in marathon varies across lifespan with a quadratic relationship between age and marathon performance. Male runners have a faster run time compared female runners across lifespan. In addition, female runners are generally slower in all marathon compared to males runners, especially from age 40 and above. male runners have a faster run time compared female runners across lifespan. We can see that the estimated line of best fit for gender is similar at earlier stages, but deviates more as age increases. The corresponding table (Table 3) below shows the best runner from the marathon data set is a 25 year old male from Chicago with a percent off course record of -0.78 and run time of 135.9276 minutes. All the best runners from the marathon data set are also male. We can also see that the fastest runner in Grandma's marathon is the slowest compared to the fastest runner in other marathons. This could be due to the fact that Grandma's marathon has the highest WBGT among all marathon locations.

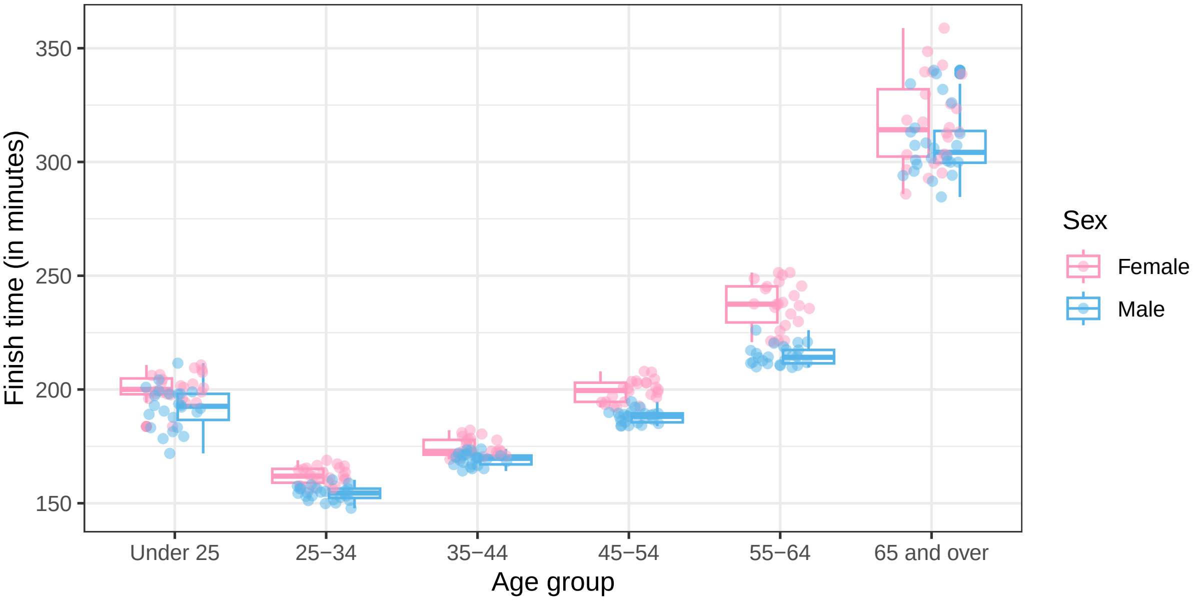
Table 3: Top runner in each marathon

Marathon	Year	Sex	Age	% off Course Record	Finish time (mins)	WBGT	Flag
Chicago	2012	Male	25	-0.7828048	135.9276	6.71000	Green
Boston	2011	Male	29	-2.2510593	136.8485	6.96500	Green
New York City	2011	Male	30	-2.0360219	139.1088	NA	Black
Twin Cities	2016	Male	28	-0.9481102	144.6158	12.00167	Red
Grandma's	2014	Male	26	-0.3986113	145.4180	18.04444	White

The boxplot (Figure 3) shows the distribution of actual course record by age group category. I categorized the age group into six categories. The boxplot shows that the fastest runners are between 25-34 years old. This boxplot shows the same trend as the scatter plot in Figure 2 that there's negative effect of age on marathon performance after their peak performance at age 30. The boxplot also provides evidence that female runners are slower than male runners in all age groups, and the magnitude of the difference in marathon performace is larger in the older age groups between men and women, especially in the 55-64 and 65 and over age groups. We can observe that the distributions of male and female actual course records have some overlap in age groups "Under 25", "25-34", and "35-44". However, the gender distributions differ in age groups "45-54" ,"55-64", and "65 and over", suggesting that women show accelerated decline in marathon performance and tend to be slower as they age compared to men.

Figure 3: Course record by age group category

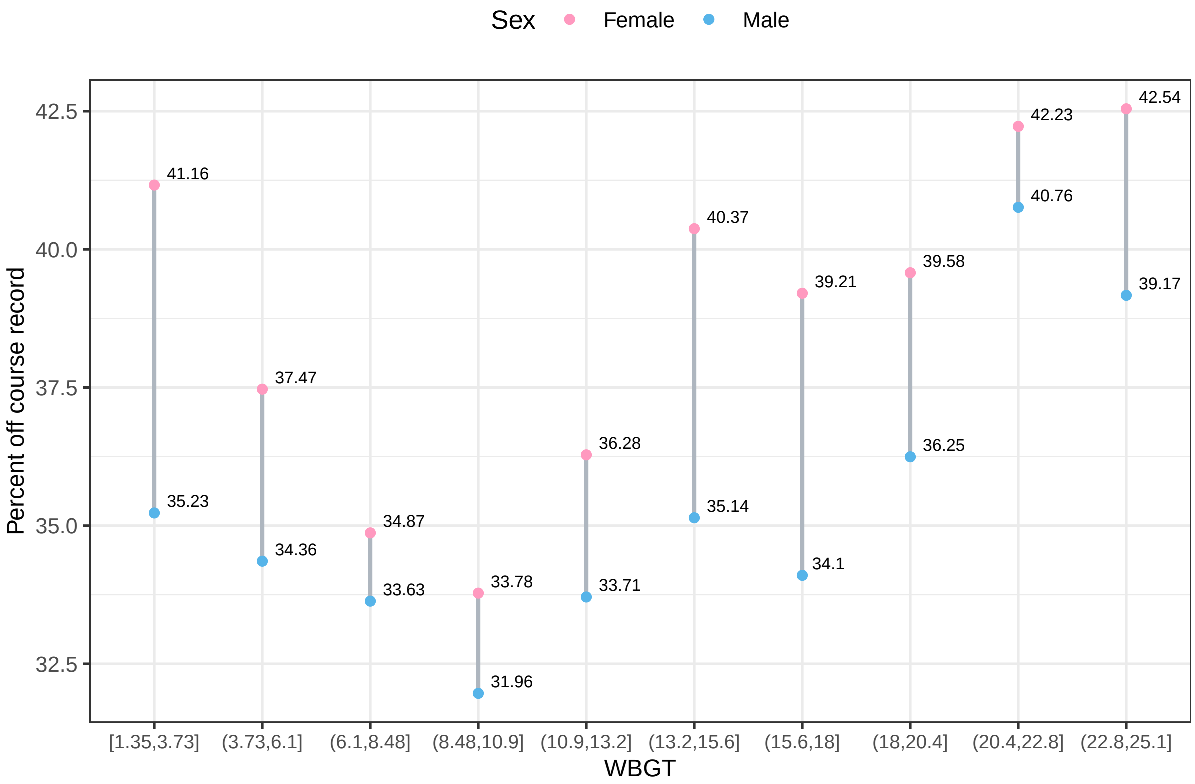
Comparison for course record distribution among age groups and gender



Aim 2: Explore the impact of environmental conditions on marathon performance, and whether the impact differs across age and gender.

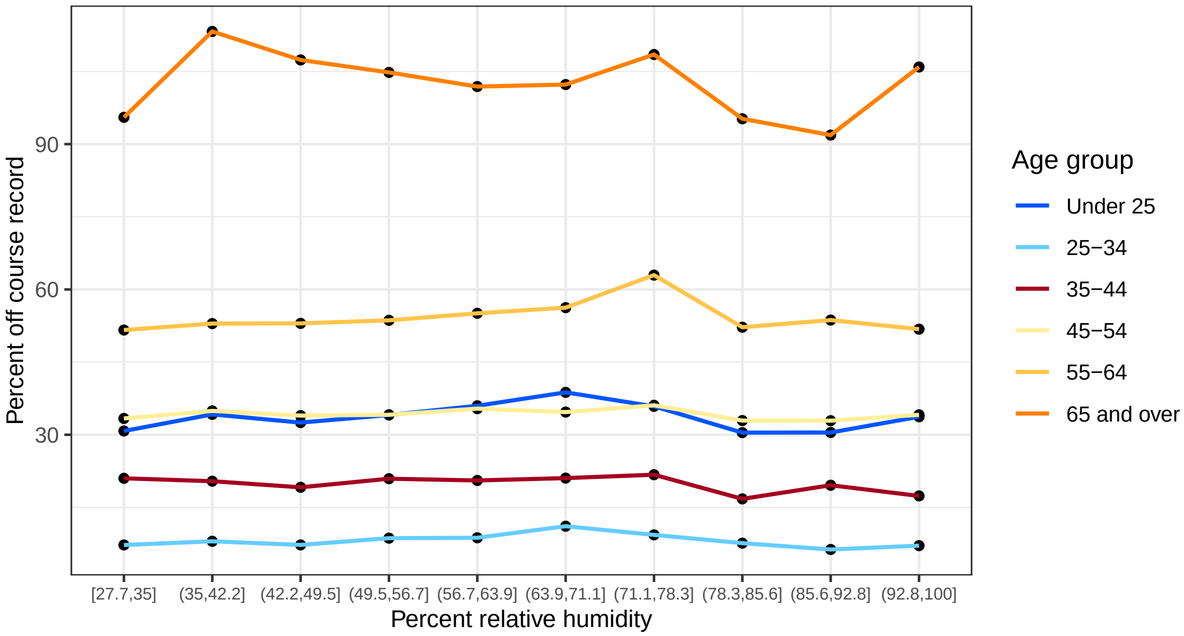
I focused on the wet bulb globe temperature (WBGT) for examining the impact of environmental conditions because it is a parameter calculated using 3 temperatures (dry bulb, black globe, and wet bulb). I summarized the WBGT into 10 intervals and calculated the median of percent off course record for each WBGT interval. The intervals are equally spaced and the minimum WBGT is 1.35 and the maximum WBGT is 25.1. The scatter plot (Figure 4) shows that the percent off course record is the lowest for WBGT interval (8.48,10.9] for both men and women. We can also observe that the relationship between WBGT and marathon performance is U-shaped, and the percent off course record is higher for female runners. The difference between their performance is approximately 2-5 minutes, depending on the range of the wet bulb globe temperature. Therefore, the impact of environmental condition on marathon performance is different based on the temperature level, and the impact differs between men and women.

Figure 4: Association between WBGT and marathon performance by gender



I further examined all the environmental conditions and its impact on marathon performance across age. I visualized percent relative humidity, and its impact on marathon performance in Figure 5. I summarized the percent relative humidity into 8 intervals and calculated the median of percent off course record for each humidity interval. Although, the scatter plot shows the lines for age groups are relatively parallel to each other; but, the changes in performance across humidity is some what different between age groups. We can observe that younger age groups such as “25-34”, and “35-44” doesn’t show much difference in performance across humidity levels. However, young adults “Under 25” and older age groups such as “55-64” and “65 and over” show more deviation in performance at different levels of relative humidity. Therefore, this entails that certain age groups are more sensitive to heat and humidity, and the impact of environmental conditions on marathon performance differs across age groups.

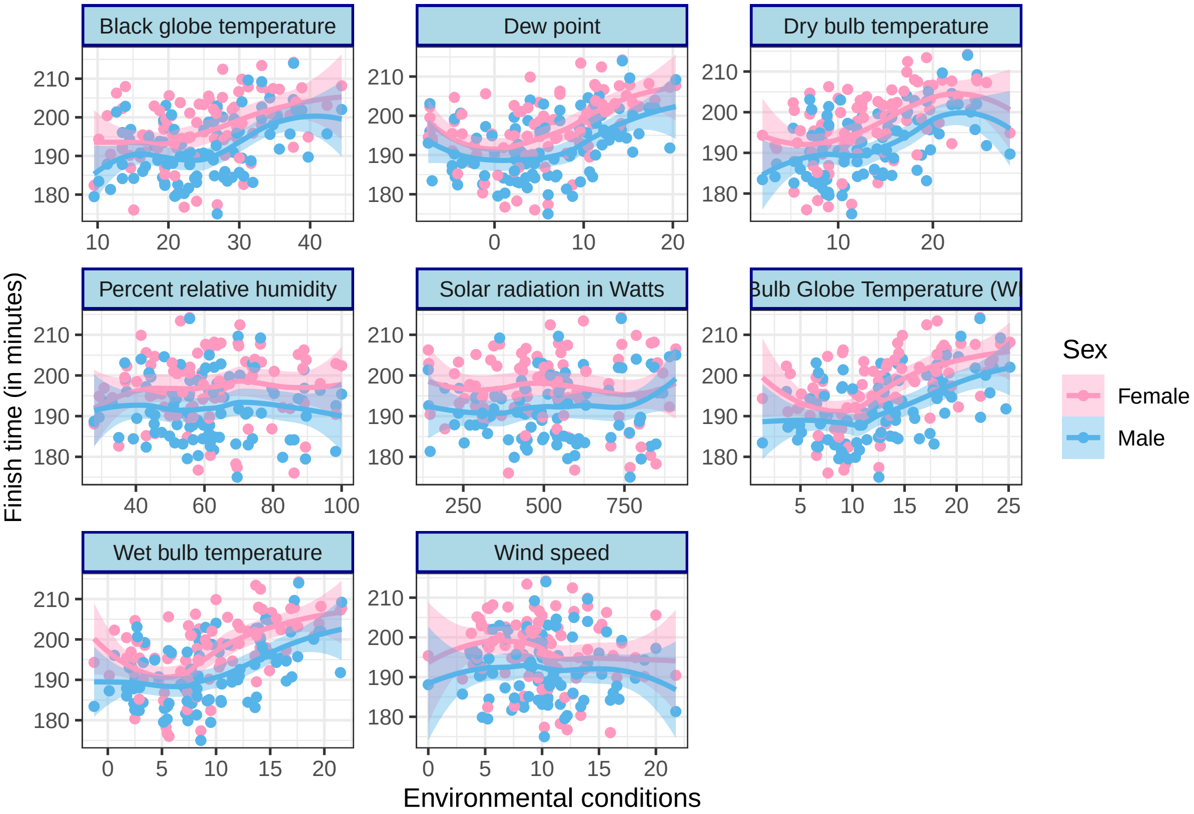
Figure 5: Association between percent relative humidity and marathon performance



Aim 3: Identify the weather parameters (WBGT, Flag conditions, temperature, etc) that have the largest impact on marathon performance

I estimated the median finish time of each marathon race for both men and women and plotted against all the weather parameters in the data set to get a holistic view of the weather impact on marathon performance. It seems like black globe temperature, dry bulb temperature, wet bulb temperature, wet bulb globe temperature, and dew point have a negative impact on marathon performance as the temperature rises (see Figure 6). The fitted local polynomial regression line shows that lower temperature is associated with faster marathon performance, and as the temperature increases, the marathon performance decreases. The scatter plot also shows that the impact of environmental conditions on marathon performance different for male and female, because female runners have longer finish time for higher temperatures compared to male runners across weather parameters such as black globe temperature, dry bulb temperature, wet bulb temperature, and dew point. Environmental conditions such as percent relative humidity, solar radiation and wind speed have less impact on marathon performance as they are less correlated with marathon performance.

Figure 6: Association between environmental conditions and marathon performance



Limitation

The data set only includes top single-age performances; therefore, it does not include all runners who participated in the marathons. As a result, the data may not be representative of the entire population of marathon runners. The data set records wind speed, but it does not account for wind direction, which can either help or hinder marathon performance. Some single-age performances exhibit extreme values (> 700 minutes) for marathon finish time, which require further investigation to ensure data quality. The impact of air quality on marathon performance should also be explored. However, the air quality index data set contains only a few observations for PM2.5 (code = 88502), so it cannot be used to assess air quality's impact on marathon performance. The table below shows the summary of air quality index for marathons, and the p-value indicates that there is a significant difference in air quality index between marathon locations. Therefore, it is important to consider air quality index as a potential confounder in the analysis.

Table 4: Air quality index for marathons

AQI	**Boston** N = 19	**Chicago** N = 21	**Grandma's** N = 17	**New York City** N = 24	**Twin Cities** N = 17	**p-value**
Ozone	0.04 (0.03, 0.04)	0.02 (0.02, 0.03)	0.03 (0.03, 0.03)	0.02 (0.02, 0.02)	0.02 (0.02, 0.03)	<0.001
PM2.5	8.1 (5.7, 11.2)	10.6 (7.1, 18.6)	6.8 (4.1, 8.3)	7.3 (5.0, 11.3)	5.5 (3.5, 7.9)	0.008

[1] Median (Q1, Q3)

[2] Kruskal-Wallis rank sum test

Conclusion

In this EDA, we have discovered that marathon performance varies across lifespan in both men and women. There's a quadratic relationship between age and marathon performance, with runners peaking in their mid-20s to early 30s before experiencing a decline. Female runners are generally slower than male runners across the lifespan. This analysis also suggest that women experience worse decline in marathon performance as they age because the difference in marathon finish time widens significantly in older age groups, particularly after age 40. There's also evidence that environmental conditions such as temperature have a negative impact on marathon performance. The relationship between WBGT and marathon performance follows a U-shaped curve, with the optimal performance occurring at moderate WBGT levels (8.48-10.9) in both men and women. The impact of environmental conditions also differs among different age groups. For example, younger age group and older age group runners show more deviation in performance as humidity increases compared to middle age groups. Lastly, temperature (black globe temperature, dry bulb temperature, wet bulb temperature, and dew point) seems to be the most important factor that affects marathon performance, with moderate temperature associated with best marathon performance. Wind speed, solar radiation, and percent relative humidity have less impact on marathon performance.

References

Besson, Thibault, Robin Macchi, Jeremy Rossi, Cédric Y M Morio, Yoko Kunimasa, Caroline Nicol, Fabrice Vercruyssen, and Guillaume Y Millet. 2022. "Sex Differences in Endurance Running." *Sports Med* 52 (6): 1235–57. https://doi.org/10.1007/s40279-022-01651-w.

ELY, MATTHEW R., SAMUEL N. CHEUVRONT, WILLIAM O. ROBERTS, and SCOTT J. MONTAIN. 2007. "Impact of Weather on Marathon-Running Performance." *Medicine & Science in Sports & Exercise* 39 (3). https://journals.lww.com/acsm-msse/fulltext/2007/03000/impact_of_weather_on_marathon_running_performance.12.aspx.

Kenney, W. Larry, and Thayne A. Munce. 2003. "Invited Review: Aging and Human Temperature Regulation." *Journal of Applied Physiology* 95 (6): 2598–2603. https://doi.org/10.1152/japplphysiol.00202.2003.

Code Appendix

```r
# a few settings we define for the file
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
knitr::opts_chunk$set(warning = FALSE)
knitr::opts_chunk$set(fig.width = 8, fig.height = 6)
knitr::opts_chunk$set(tidy = TRUE, kable = TRUE)
knitr::opts_chunk$set(fig.align = "center")
# load libraries
library(tidyverse) # data manipulation
library(dplyr) # data manipulation
library(ggplot2) # visualization
library(kableExtra) # create nice table
library(gtsummary) # create table summary and statistics
library(lubridate) # use to reformat dates
theme_gtsummary_compact() #Setting theme "Compact" for gtsummary tables
# set working directory
setwd("~/Desktop/PHP 2550 Pratical Data Analysis/Project 1")
# load data
marathon <- read.csv("~/Desktop/PHP 2550 Pratical Data Analysis/Project 1/project1.csv")
course_rec <- read.csv("course_record.csv")
aqi <- read.csv("aqi_values.csv")
# check the data is in the right format and factor any categorical variables
marathon <- rename(marathon, Race = Race..0.Boston..1.Chicago..2.NYC..3.TC..4.D.,
                   Sex = Sex..0.F..1.M., Age = Age..yr.)
marathon$Race <- as.factor(marathon$Race)
marathon$Sex <- as.factor(marathon$Sex)
marathon$Flag <- as.factor(marathon$Flag)
levels(marathon$Flag) <- c("Black", "Red","Yellow","Green","White")
marathon$Year <- as.factor(marathon$Year)
# factor course record data set and relabel variable
course_rec <- course_rec %>%
  mutate(Race = as.factor(case_when(Race == "B" ~ 0, Race == "C" ~ 1, Race =="NY" ~ 2,
                          Race == "TC" ~ 3, Race == "D" ~ 4)))
course_rec$Year <- as.factor(course_rec$Year)

# merge the data sets marathon and course_rec by race and year
marathon <- left_join(marathon, course_rec, by = c("Race", "Year"))

# calculate the true finish time for each runner
marathon$CR <- hms(marathon$CR) # format to 'hours:minutes:seconds'
```

```
marathon$CR <- hour(marathon$CR)*60 + minute(marathon$CR) # convert to minutes
marathon <- marathon %>%
  mutate(actual_CR = marathon$CR + marathon$CR* (marathon$X.CR/100))

# change the values of percent relative humidity less than 1 into percent
marathon <- marathon %>%
  mutate(X.rh = ifelse(X.rh <= 1, X.rh*100, X.rh))
# relabeling, rename variables, and create a location variable
marathon <- marathon %>%
  mutate(Sex = ifelse(Sex == "0", "Female", "Male"),
         Location = as.factor(case_when(Race == "0" ~ "Boston", Race == "1" ~ "Chicago", Race
  rename(`Percent off course record` = X.CR,
         `Dry bulb temperature` = Td..C,
         `Wet bulb temperature` = Tw..C,
         `Percent relative humidity` = X.rh,
         `Wind speed` = Wind,
         `Wet Bulb Globe Temperature (WBGT)` = WBGT,
         `Black globe temperature` = Tg..C,
         `Solar radiation in Watts`= SR.W.m2,
         `Dew point`= DP)

tbl_summary(marathon,
            include = c(actual_CR,Location, Sex, Age), by = Sex, missing = "no",
            label = list(actual_CR = "Actual course record (in minutes)")) %>%
  modify_spanning_header(all_stat_cols() ~ "**Gender**") %>%
  modify_caption("Characteristics of marathon runners") %>%
  as_kable_extra(booktabs = TRUE) %>%
  kable_styling(font_size = 11, full_width = F, latex_options = c("repeat_header", "HOLD_pos
table1 <- marathon %>%
  group_by(Year, Location) %>%
  summarize(`Dry bulb temperature` = unique(`Dry bulb temperature`),
            `Wet bulb temperature` = unique(`Wet bulb temperature`),
            `Percent relative humidity` = unique(`Percent relative humidity`),
            `Wind speed` = unique(`Wind speed`),
            `Wet Bulb Globe Temperature (WBGT)` = unique(`Wet Bulb Globe Temperature (WBGT)`
            `Black globe temperature` = unique(`Black globe temperature`),
            `Solar radiation in Watts` = unique(`Solar radiation in Watts`),
            `Dew point` = unique(`Dew point`))

tbl_summary(data = table1, include = c(-Year), by = Location, missing = "no",
            label = list(`Wet Bulb Globe Temperature (WBGT)` = "WBGT",
                         `Wet bulb temperature` = "Wet bulb",
                         `Dry bulb temperature` = "Dry bulb",
                         `Black globe temperature` = "Black globe",
                         `Solar radiation in Watts` = "Solar radiation",
                         `Percent relative humidity` = "% relative humidity"
                         ),
            digits = list(everything() ~ c(0))) %>%
  add_p() %>%
  modify_header(label = "**Measurements**", stat_4 = "**NYC** \n N= {n}") %>%
  modify_spanning_header(all_stat_cols() ~ "**Marathons**")  %>%
  modify_caption("Environmental conditions for marathon locations") %>%
  as_kable_extra(booktabs = TRUE) %>%
  kable_styling(latex_options = c("repeat_header", "HOLD_position"), font_size = 10, full_wi
# create heatmap for environmental conditions across all marathon locations and years

# define colors for WBGT
cols <- c("white",
             "honeydew1", #lightblue
             "#4ab04a", #green
             "#ffd73e", #yellow
             "#ce472e", #red
             "#000000") #black
marathon %>%
  select(Location, Year, `Dry bulb temperature`, `Wet bulb temperature`, `Percent relative hu
  pivot_longer(cols = -c(Location, Year), names_to = "Environmental conditions", values_to =
  filter(`Environmental conditions` == "Wet Bulb Globe Temperature (WBGT)") %>%
  ggplot(aes(x = Year, y = Location, fill = Values)) +
  geom_tile(color="white", size = 0.35) +
  scale_fill_gradientn(name = "WBGT", breaks = c(0, 10, 15, 18, 23, 28),
                       labels = c("0","10", "10-18", "18-23", "23-28", "28"),
                       colors = cols, na.value = "white") +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5, size = 12), panel.grid = element_blank(),
        axis.text.x = element_text(angle = 45, hjust=1)) +
  labs(
    x = "",
    y = "",
    fill = "Values",
    title = "Figure 1: Heatmap of WBGT across marathons and years"
  ) +
  scale_x_discrete(expand = c(0,0.5))
## Correlation plot - NOT INCLUDED IN THE FINAL REPORT
numeric_data <- marathon %>%
  select(`Percent off course record`, `Dry bulb temperature`, `Wet bulb temperature`, `Percen

cor_matrix <- cor(numeric_data, use = "complete.obs")
corrplot(cor_matrix, method = "color", type = "full", tl.col = "black", tl.cex = 0.7, cl.cex
# Scatterplot for fastest or best CR vs. age for men vs. women
new_labels <- c("0" = "Boston", "1" = "Chicago", "2" = "New York City", "3" = "Twin Cities",

graph1 <- marathon %>%
  group_by(Race, Age, Sex) %>%
  summarize(actual_CR = mean(actual_CR))
#FF99BF
#f2bbfc
cols <- c(Female = "#FF99BF", Male = "#56B4E9")
ggplot(data = graph1, aes(y = actual_CR, x = Age, group = Sex, Color = Sex)) +
  geom_point(aes(color = Sex)) +
  geom_smooth(aes(colour = Sex, fill = Sex)) +
  theme_bw() +
  facet_wrap(~Race, scales = "free", labeller = labeller(Race = new_labels)) +
  scale_color_manual(name = "Sex",breaks = c("Female", "Male"), values = cols) +
  scale_fill_manual(name = "Sex",breaks = c("Female", "Male"), values = cols) +
  labs(
    x = "Age",
    y = "Runner's finish time (in minutes)",
    title = "Figure 2: Association of course record with age and gender",
    subtitle = ""
  ) +
  theme(strip.background = element_rect(fill="lightblue", size=1, color="darkblue"),
        axis.title.y = element_text(size = 10),
        plot.title = element_text(size = 12, hjust = 0.5),
        legend.position = "bottom"
        )
marathon %>%
  group_by(Location) %>%
  slice_min(order_by = actual_CR, n = 1) %>%
  select(Location, Year, Sex, Age, `Percent off course record`, actual_CR, `Wet Bulb Globe Te
  rename(Marathon = Location, `Finish time (mins)` = actual_CR,
         "% off Course Record" = `Percent off course record`,
         "WBGT" = `Wet Bulb Globe Temperature (WBGT)`) %>%
  arrange(`Finish time (mins)`) %>%
  as.data.frame() %>%
  kable(format = "latex", caption = "Top runner in each marathon", booktab = T, linesep = "")
  kable_styling(latex_options = c("repeat_header", "HOLD_position"), font_size = 10, position
# Examine aging on marathon performance
# Create a new variable for age group
marathon <- marathon %>%
  mutate(Age_group = case_when(Age < 25 ~ "Under 25",
                               Age >= 25 & Age < 35 ~ "25-34",
                               Age >= 35 & Age < 45 ~ "35-44",
                               Age >= 45 & Age < 55 ~ "45-54",
                               Age >= 55 & Age < 65 ~ "55-64",
                               Age >= 65 ~ "65 and over"))
marathon$Age_group <- factor(marathon$Age_group, levels = c("Under 25", "25-34", "35-44", "45
marathon %>%
  group_by(Age_group, Sex, Year) %>%
  summarize(actual_CR = mean(actual_CR)) %>%
  ggplot(aes(x = Age_group, y = actual_CR, color = Sex)) +
  geom_boxplot() +
  scale_color_manual(name = "Sex",breaks = c("Female", "Male"), values = cols) +
  scale_fill_manual(name = "Sex",breaks = c("Female", "Male"), values = cols) +
  geom_jitter(width = 0.2, height = 0, alpha = 0.5) +
  theme_bw() +
  labs(y = "Finish time (in minutes)", x = "Age group", title = "Figure 3: Course record by a
       subtitle = "Comparison for course record distribution among age groups and gender") +
  theme(plot.title = element_text(hjust = 0.5),
        plot.subtitle = element_text(hjust = 0.5))
# Analysis for WBGT impact on marathon performance
seq_wbgt <- seq(min(marathon$`Wet Bulb Globe Temperature (WBGT)`, na.rm = T), max(marathon$`W

WBGT <- marathon %>%
  mutate(WBGT_intervals = cut(`Wet Bulb Globe Temperature (WBGT)`, breaks = seq_wbgt, include
  filter(!is.na(`Wet Bulb Globe Temperature (WBGT)`)) %>%
  group_by(WBGT_intervals, Sex) %>%
  summarize(across(`Percent off course record`:`actual_CR`, median, na.rm = TRUE),
            n = n())

# Percent off course record by WBGT intervals
Males <- WBGT %>%
  filter(!is.na(`Wet Bulb Globe Temperature (WBGT)`)) %>%
  group_by(WBGT_intervals, Sex) %>%
  summarize(across(`Percent off course record`:`actual_CR`, median, na.rm = TRUE)) %>%
  filter(Sex == "Male")

Females <- WBGT %>%
  filter(!is.na(`Wet Bulb Globe Temperature (WBGT)`)) %>%
  group_by(WBGT_intervals, Sex) %>%
  summarize(across(`Percent off course record`:`actual_CR`, median, na.rm = TRUE)) %>%
  filter(Sex == "Female")

gender_combined <- WBGT %>%
  filter(!is.na(`Wet Bulb Globe Temperature (WBGT)`)) %>%
  group_by(WBGT_intervals, Sex) %>%
  summarize(across(`Percent off course record`:`actual_CR`, median, na.rm = TRUE))

ggplot(gender_combined) +
  geom_segment(data = Males, aes(x = WBGT_intervals, xend = Females$WBGT_intervals, yend = Fe
  geom_point(aes(WBGT_intervals, y = `Percent off course record`, color = Sex), size = 1.5) +
 geom_text(aes(WBGT_intervals, y = `Percent off course record`, label = round(`Percent off co
  theme_bw() +
  scale_color_manual(name = "Sex",breaks = c("Female", "Male"), values = cols) +
  ggtitle("Figure 4: Association between WBGT and marathon performance by gender") +
  labs(x = "WBGT", y = "Percent off course record") +
  theme(plot.title = element_text(hjust = 0.5),
        axis.title.y = element_text(size = 10),
        axis.title.x = element_text(size = 10),
        axis.text.x = element_text(size = 8),
        legend.position = "top"
  )
# define colors for age group
cols_age <- c("Under 25" = "#0055FF", "25-34" = "#66CCFF", "35-44" = "#A50021", "45-54" = "#

# Percent relative humidity intervals
marathon <- marathon %>%
  mutate(Humidity_intervals = cut(`Percent relative humidity`, breaks = 10, include.lowest =

humidity <- marathon %>%
  filter(!is.na(Humidity_intervals)) %>%
  group_by(Humidity_intervals, Age_group) %>%
  summarize(across(`Percent off course record`:`actual_CR`, median, na.rm = TRUE))

humidity_plot <- ggplot(data = humidity) +
  geom_point(aes(x = Humidity_intervals, y = `Percent off course record`), size = 1.5) +
 geom_line(aes(x = Humidity_intervals, y = `Percent off course record`, color = Age_group, g
  theme_bw() +
  scale_color_manual(name = "Age group", values = cols_age) +
  ggtitle("Figure 5: Association between percent relative humidity and marathon performance")
  labs(x = "Percent relative humidity", y = "Percent off course record") +
  theme(plot.title = element_text(hjust = 0.5, size = 12),
        axis.title.y = element_text(size = 10),
        axis.title.x = element_text(size = 10),
        axis.text.x = element_text(size = 7)
  )
humidity_plot
# Examine the association between environmental conditions and marathon performance
# Take the median of environmental conditions and course record for each race
marathon %>%
  group_by(Year, Location, Sex) %>%
  summarize(across(`Percent off course record`:`actual_CR`, median, na.rm = TRUE)) %>%
  select(Sex,`actual_CR`, `Dry bulb temperature`, `Wet bulb temperature`, `Percent relative
  pivot_longer(cols = -c(`actual_CR`, Year, Sex, Location),
               names_to = "Environmental conditions", values_to = "Values") %>%
  ggplot(aes(y = `actual_CR`, x = Values, color = Sex, fill = Sex)) +
  geom_point() +
  geom_smooth(se = TRUE) +
  facet_wrap(~`Environmental conditions`, scales = "free") +
  theme_bw() +
  scale_color_manual(name = "Sex",breaks = c("Female", "Male"), values = cols) +
  scale_fill_manual(name = "Sex",breaks = c("Female", "Male"), values = cols) +
  labs(
    x = "Environmental conditions",
    y = "Finish time (in minutes)",
    title = "Figure 6: Association between environmental conditions and marathon performance"
  theme(strip.background = element_rect(fill="lightblue", size=1, color="darkblue"),
        axis.title.y = element_text(size = 10),
        plot.title = element_text(size = 12, hjust = 0.5)
        )
# reformat dates into years and relabel location
aqi$yrs <- year(as.Date(aqi$date_local,format = "%m/%d/%Y"))
aqi$Year <- as.factor(ifelse(aqi$yrs>90, aqi$yrs+1900, aqi$yrs+2000))
aqi <- aqi %>%
  mutate(Location = as.factor(case_when(marathon == "NYC" ~ "New York City",
                                       marathon == "Grandmas" ~ "Grandma's",
                                        TRUE ~ marathon)))

# change each parameter code into a column
aqi_mean <- aqi %>%
  group_by(Year, Location,parameter_code) %>%
  summarise(arithmetic_mean = mean(arithmetic_mean)) %>%
  pivot_wider(names_from = parameter_code, values_from = arithmetic_mean) %>%
  select(-`88502`) %>%
  rename(Ozone = `44201`, PM2.5 = `88101`)

# merge marathon data with AQI data set by marathon
marathon <- left_join(marathon, aqi_mean, by = c("Location", "Year"))

# create a summary table for AQI
aqi_mean %>%
  tbl_summary(include = c(Location, `Ozone`, `PM2.5`), by = Location, missing = "no",
              digits = list(Ozone ~ c(2))) %>%
  add_p() %>%
  modify_caption("Air quality index for marathons") %>%
  modify_header(label = "**AQI**") %>%
  as_kable_extra(booktabs = TRUE) %>%
  kable_styling(latex_options = c("repeat_header", "HOLD_position"), font_size = 10, full_wid
```