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

December 13 2024

Miaoyan Chen

Introduction

The objective of this exploratory data analysis is to investigate the impact of weather on marathon performance across age in both men and women. Preliminary findings suggest that marathon performance deteriorates as environmental temperatures increase, with this effect being more significant in longer-distance events (Ely et al. 2007). Additionally, older runners exhibit greater sensitivity to heat stress compared to their younger counterparts (Kenney and Munce 2003). Furthermore, there are well-documented differences in endurance performance between genders (Besson et al. 2022). In this study, we aim to explore the specific effects of environmental conditions—such as temperature, humidity, solar radiation, and wind—on marathon performance, with a focus on identifying variation in effects between men and women. We hypothesized that performance slowdowns would be more significant under high Wet Bulb Globe Temperature (WBGT) conditions, particularly among older individuals compared to younger ones. It is anticipated that similar trends will be observed across both genders.

To address these objectives, we utilized a comprehensive marathon dataset to conduct a series of exploratory analyses. These analyses examine how increasing age affects marathon performance in men and women, assesse the role of environmental conditions, and determine whether these weather impacts differ by age and gender. Finally, we aim to identify the weather parameters with the greatest influence on marathon performance. The analysis commences with a thorough data quality assessment, including the identification of missing data patterns. This preliminary step ensures the reliability of the dataset before conducting the exploratory analyses.

Data quality, cleaning and missingness

The marathon data set contains 11564 records and 14 parameters. Among these, race, year, sex, and flag are recorded as categorical variables, while age and weather-related 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 (WBGT), are continuous variables. The dataset includes top single-age performances from five major marathons—Boston, Chicago, New York City (NYC), Twin Cities, and Grandma's (Duluth)—spanning the years 1993 to 2016 for both men ($n = 6,112$) and women ($n = 5,452$). Additionally, detailed weather and environmental conditions for each race are recorded within the dataset.

Overall, the dataset is well-structured, and a comprehensive data codebook provides clear explanations of each variable. As part of the data cleaning process, categorical variables were converted into factor variables, and all weather parameters were ensured to be numerically coded. Variables were renamed to enhance the clarity of tables and figures. The data set was then merged with the course record dataset to calculate the actual course record (finish time) for each runner. This calculation involved adding the percent off course record to the best course record for each runner. The distribution of continuous variables was examined, revealing that these environmental conditions were not normally distributed. Consequently, the median was used to summarize the percent off course record and runner's finish time in the analysis. Percent relative humidity is skewed, with many observations less than or equal to 1, likely due to a data entry error. To address this issue, values less than 1 were converted into percentages for analysis.

The only missing data in this dataset pertains to observations from the 2011 and 2012 marathon races. Specifically, there are no records for the Chicago, NYC, Twin Cities, or Grandma's marathons in 2011. This missingness appears to be completely random (MCAR) and is likely attributable to the cancellation of these races or the unavailability of weather data during those years. Since the missingness is unrelated to data entry errors, it is not addressed further in the analysis.

Another set of data, the Air Quality Indec (AQI) data set conatains different sampling locations and sampling durations for marathon air qualities. The AQI data set is also merged with the marathon data set by marathon location and date to obtain air quality information for each race. There are two sampling durations for AQI data set - 1 hour and 8 hours. Due to missingness of the 1-hour sampling duration, only 8 hour sampling duration is presented in this report.

Characteristics of marathon locations

Table 1 summarizes the environmental conditions for each marathon location across all years. The table reveals variation in weather parameters among the marathon locations. Notably, Grandma's Marathon exhibits the highest median values for dry bulb temperature, wet bulb temperature, Wet Bulb Globe Temperature (WBGT), black globe temperature, and dew point, as well as the highest levels of solar radiation. These findings suggest that Grandma's Marathon likely experiences the highest heat stress among all the marathon locations. In contrast, the Boston Marathon displays the lowest median values for dry bulb temperature, wet bulb temperature, and solar radiation.Boston and New York City share relatively similar weather conditions as they are both located in the east coast.

Table 1: 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 accompanying heatmap (Figure 1) utilizes the levels of the Flag variable, with warmer colors representing higher temperatures. The figure highlights that WBGT is highest at Grandma's Marathon, followed by the Twin Cities, Chicago, New York City, and Boston. Marathons located on the East Coast, such as those in New York City and Boston, show lower WBGT levels compared to those in the Midwest. This suggests that there's significant differences in weather and environmental conditions across marathon locations.

Figure 2 presents the correlation between the actual course record (in minutes), age, and weather parameters. We use a threshold of ≥ 0.7 to define the high correlation between variables. The figure shows that the percent off course record is positively correlated with age. This indicates that as age increases marathon performance decreases. The correlation plot also provides a comprehensive overview of the relationship weather parameters. Many of the weather parameters are highly correlated with each other, such as dry bulb temperature, wet bulb temperature, and dew point.

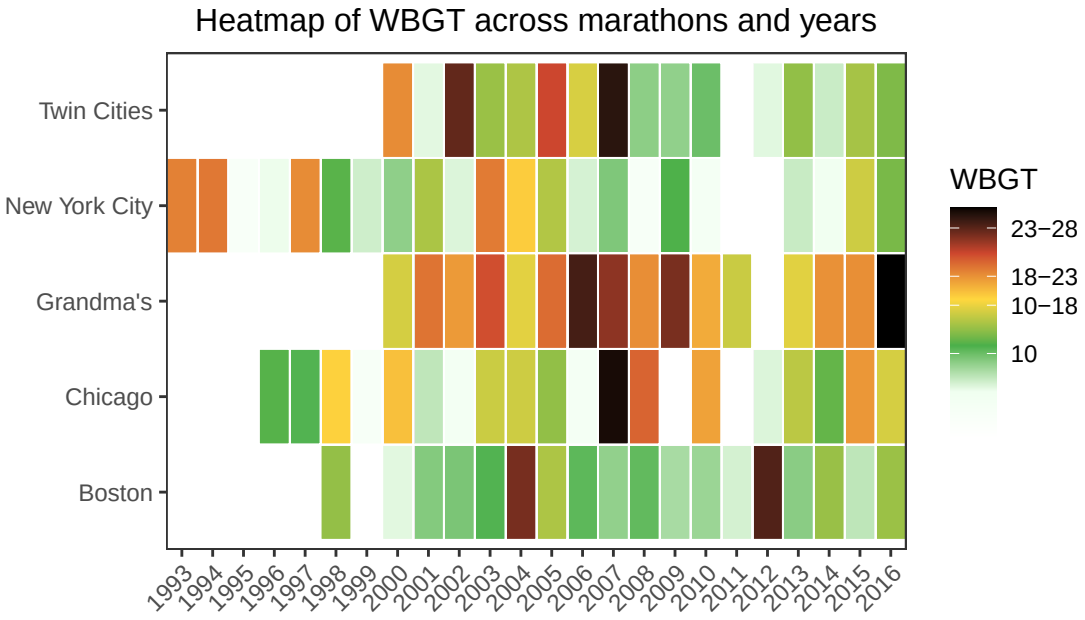

Figure 1: Heatmap of WBGT across marathons and years

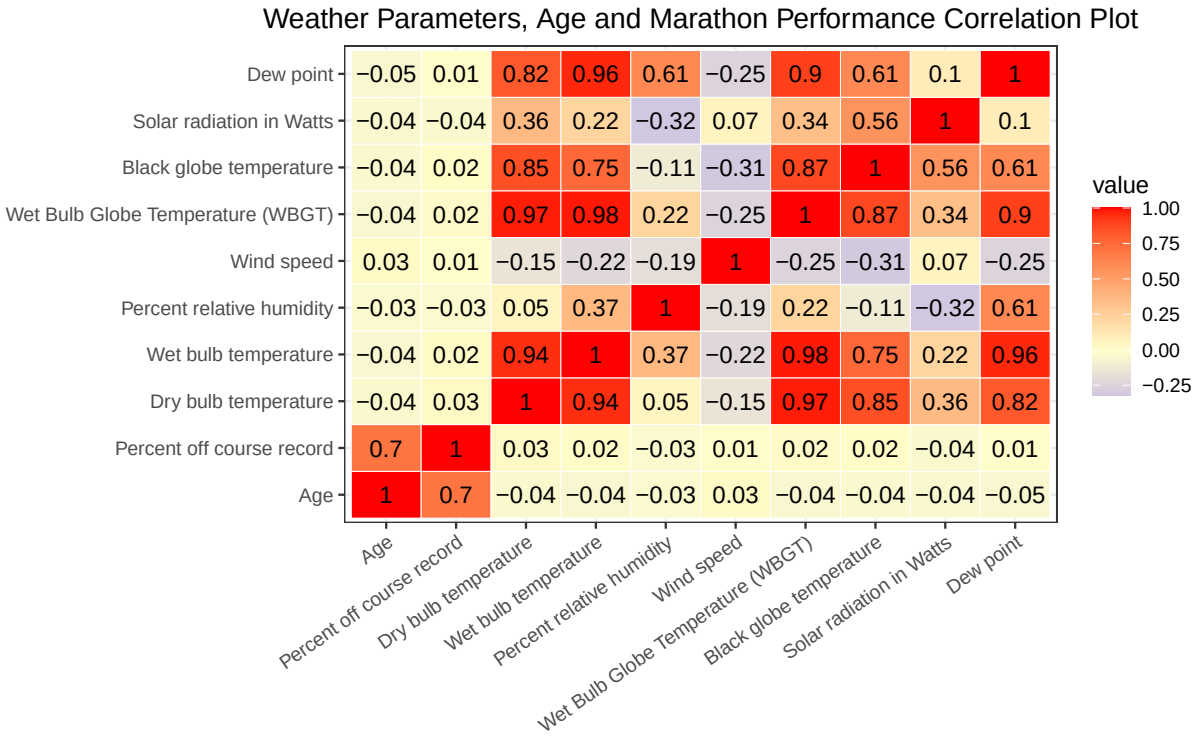

Figure 2: Weather Parameters and Marathon Performance Correlation Plot

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

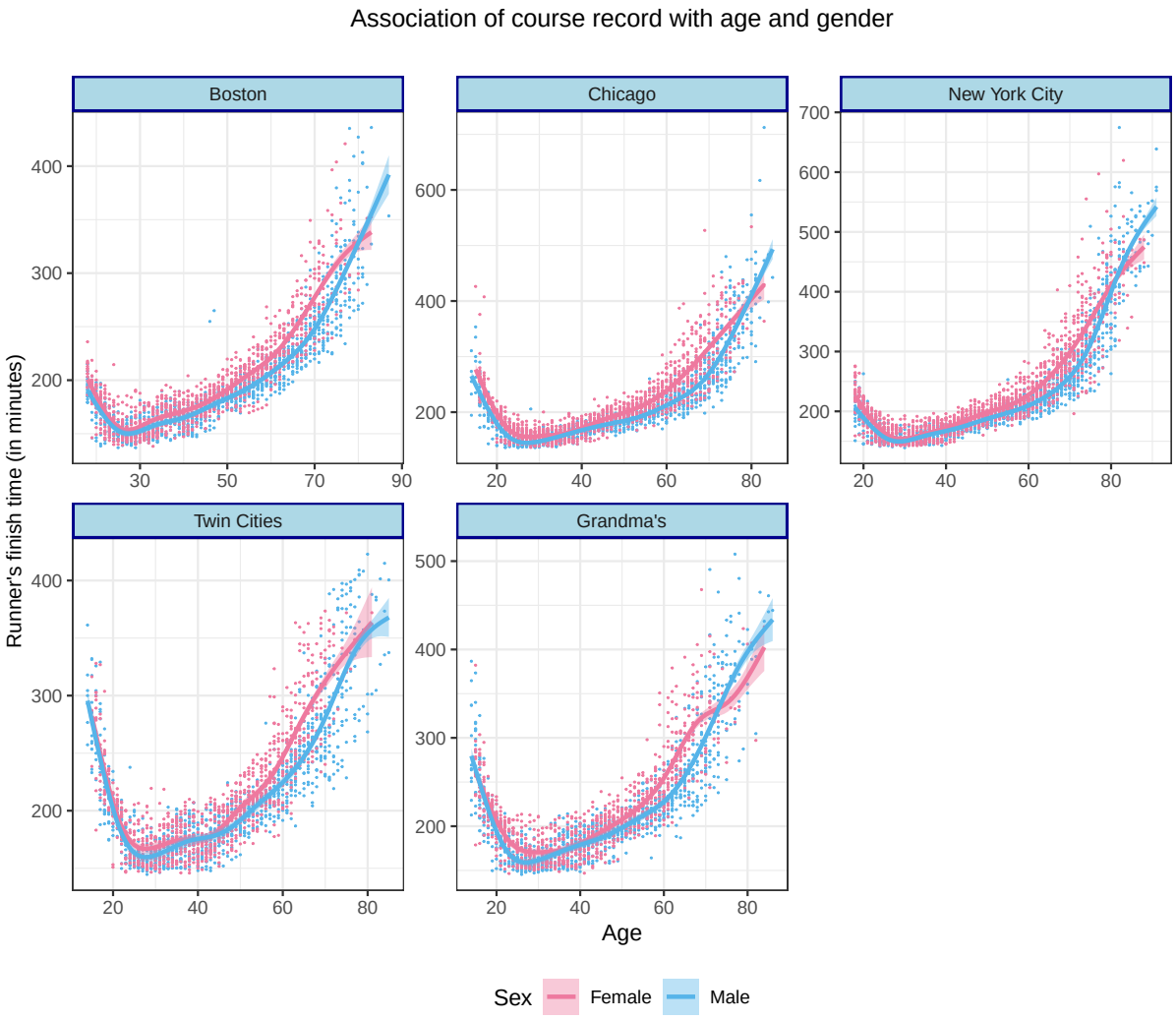

Figure 3: Association of course record with age and gender

The marathon dataset was merged with the course record dataset to calculate the actual marathon finish time for each runner. The actual course record for each runner was calculated by adding on the percentage off course record (% CR) to the best course record for each race. The course record represents the fastest time a runner has completed a marathon, and therefore a lower the course record coresponds to better performance. The scatter plot (Figure 3) illustrates the association between the course record (in minutes) and age for male and female runner at each marathon location. The relationship between runners' finish times and age exhibits a U-shaped non-linear curve. The curve shows that the fastest runners at each marathon location tend to be approximately 30 years old, after which a consistent decrease in performance is observed. Between ages 14 -30, teenage and young adult runners runners are

showing better performance as they age. Performance appears to peak around age 30, representing the optimal age for marathon success. An inflection point is present beyond age 30, performance gradually declines with increasing age.

For individuals aged approximately 20 to 60 years, an initial linear trend is observed, with better course in younger ages. However, starting from ages 60 to 70, another inflection point is evident, with finish times increasing significantly, indicating a decline in performance. This pattern underscores the significant influence of age on marathon outcomes, with distinct effects across different age groups.

Additionally, Figure 3 illustrates that male runners have a faster run time compared female runners across lifespan. Female runners are generally slower in all marathon compared to males runners, especially from age 40 and above. At younger ages, the estimated lines of best fit for male and female runners are relatively similar; however, these lines diverge significantly as age increases, highlighting a growing performance gap with age.

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

The Wet Bulb Globe Temperature (WBGT) was selected as the primary environmental parameter for examining the impact of environmental conditions on marathon performance because it is a composite measure calculated using 3 temperature variables (dry bulb, black globe, and wet bulb). To analyze its effect, WBGT was divided into 10 equally spaced intervals, ranging from a minimum of 1.35 to a maximum of 25.1. The median percent off course record was calculated for each WBGT interval. The scatter plot (Figure 4) shows that the percent off course record is lowest for WBGT intervals between 8.48 and 10.9 for both male and female runners. The relationship between WBGT and marathon performance is U-shaped, with the optimal performance occurring at moderate WBGT levels, and lower performance at both low and high WBGT levels. Furthermore, the percent off course record is consistently higher for female runners compared to male runners across all WBGT intervals, with performance differences ranging from approximately 2 to 5 minutes depending on the range of the wet bulb globe temperature. Therefore, these findings suggest that the impact of environmental condition on marathon performance is different based on the temperature level, and the impact varies between men and women.

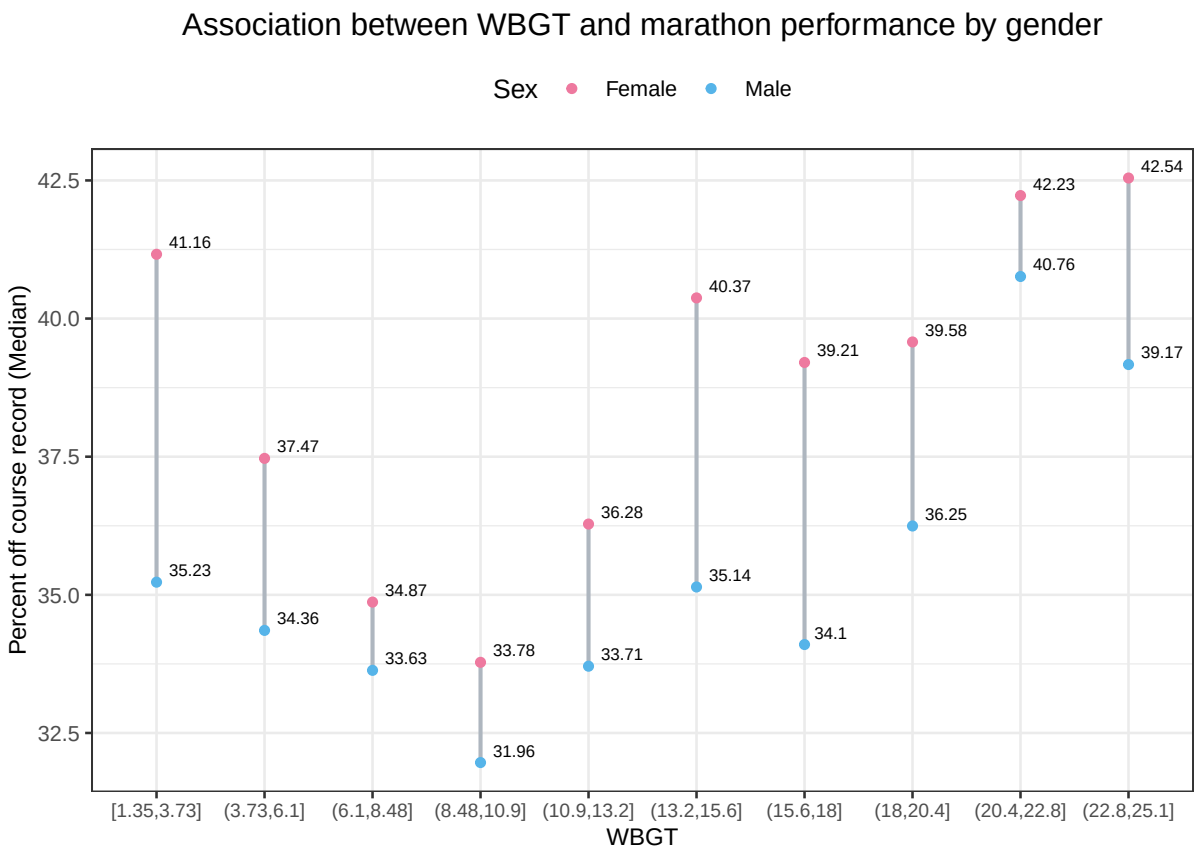

Figure 4: Association between WBGT and marathon performance

We further examined the influence of environmental conditions on marathon performance, focusing on percent relative humidity and its impact across different age groups. Percent relative humidity was divided into eight intervals, and the median percent off course record was calculated for each interval. Figure 4 visualizes the relationship between relative humidity and marathon performance. The lines representing age groups are relatively parallel across humidity levels; however, the changes in performance across humidity vary between age groups. Younger age groups, such as those aged 25–34 and 35–44, display minimal differences 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. These findings suggest that certain age groups are more sensitive to heat and humidity, and the impact of environmental conditions on marathon performance differs across age groups.

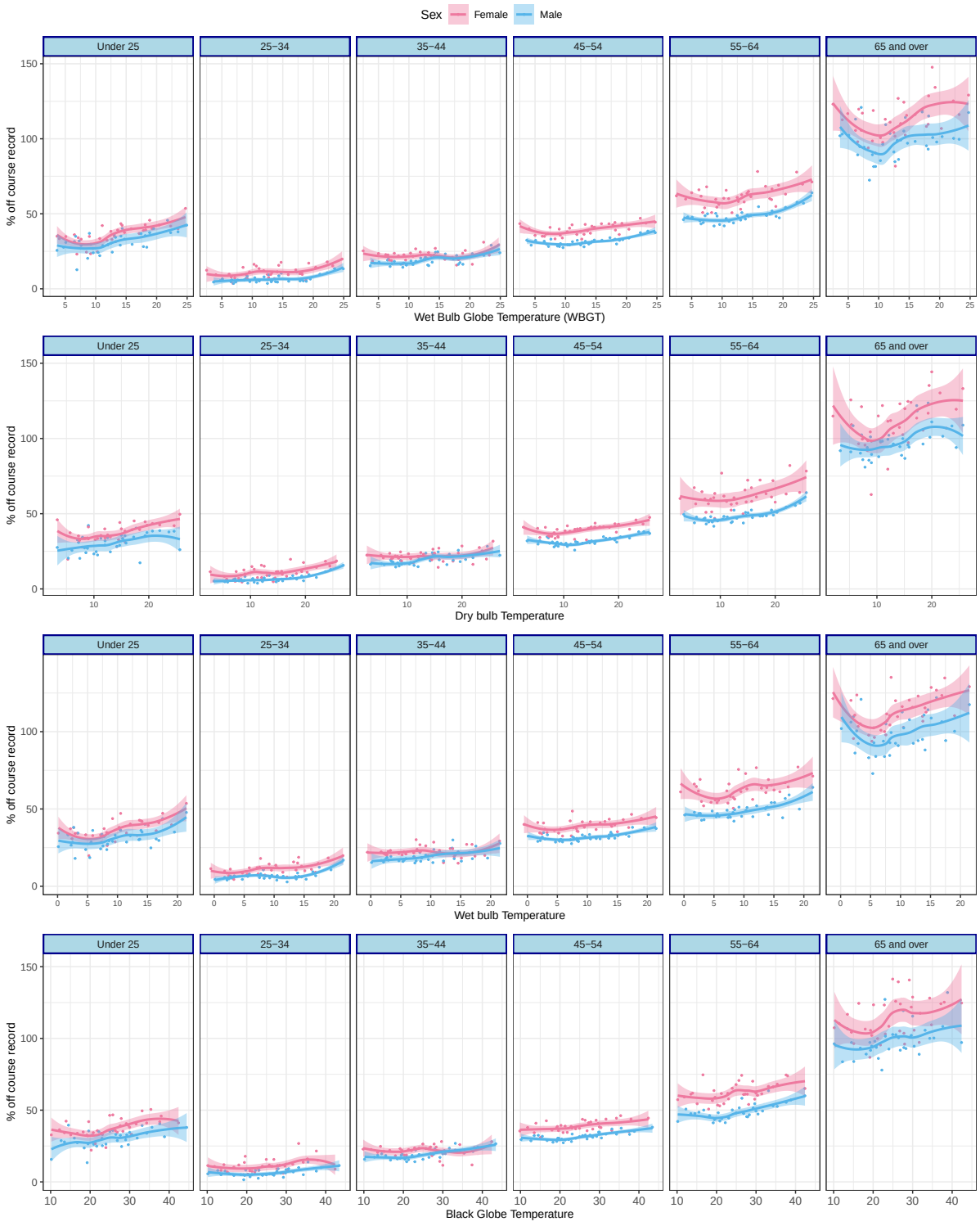

Figure 5: % off course record by temperature, sex, and age group

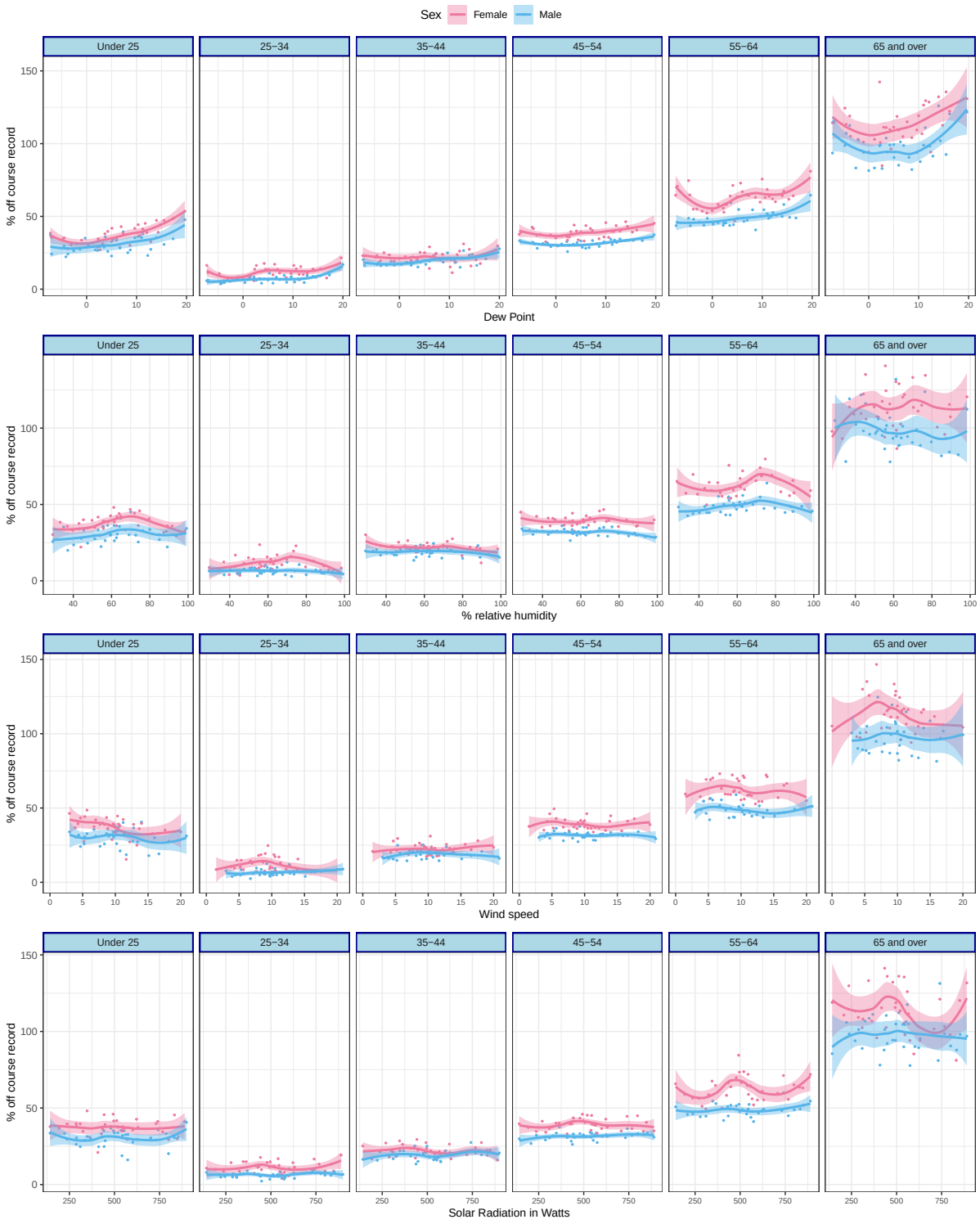

Figure 6: % off course record by environmental conditions, sex, and age group

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

To simplify the visualization and reduce the number of data points on the graph, observations were binned, with each bin represented by its median value. In Figure 5 and Figure 6, marathon performance for each marathon race for both men and women was estimated using the median percent off course record and plotted these against different weather parameters by age groups in the dataset to gain a holistic view of the impact of environmental conditions on marathon performance by sex and age. The analysis demonstrates that weather parameters such as black globe temperature, dry bulb temperature, wet bulb temperature, Wet Bulb Globe Temperature (WBGT), and dew point have most significant negative effect on marathon performance as temperatures increase. The fitted local polynomial regression lines show that lower temperatures are associated with faster marathon performance (low % off course record), while higher temperatures are associated with slower performances (high % off course record). The scatter plot further reveals that the impact of environmental conditions on marathon performance varies across sex and age groups. Older runners experience a greater decline and fluctuation in performance under extreme environmental conditions compared to younger runners. Additionally, female runners consistently exhibit longer finish times at higher temperatures compared to male runners, particularly under weather parameters such as black globe temperature, dry bulb temperature, wet bulb temperature, and dew point. The positive relationship is evident between % off course record and these weather parameters. Environmental conditions such as percent relative humidity, solar radiation and wind speed have less pronounced impact on marathon performance as they show weaker correlations with 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 2: 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, Mathew 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

```
# 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 = 7, fig.height = 5)
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(reshape2)
library(ggpubr) # arrange multiple plots
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/pro
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
                   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$Year)
# factor course record data set and relabel variable
course_rec <- course_rec %>%
  mutate(Race = as.factor(case_when(Race == "B" ~ 0, Race == "C" ~ 1, Race ==
                            Race == "TC" ~ 3, Race == "D" ~ 4)))
course_rec$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$CR) # convert to minute
marathon <- marathon %>%
  mutate(actual_CR = marathon$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" ~
  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)

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 Temp
            `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_siz
# create heatmap for environmental conditions across all marathon locations a

# 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`, `Per
  pivot_longer(cols = -c(Location, Year), names_to = "Environmental condition
  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 = eleme
        axis.text.x = element_text(angle = 45, hjust=1)) +
  labs(
    x = "",
    y = "",
    fill = "Values",
    title = "Heatmap of WBGT across marathons and years"
  ) +
  scale_x_discrete(expand = c(0,0.5))
## Correlation plot
numeric_data <- marathon %>%
  select(`Age`, `Percent off course record`, `Dry bulb temperature`, `Wet bul

cor_matrix <- cor(numeric_data, use = "complete.obs")
# corrplot(cor_matrix, method = "color", type = "full", tl.col = "black", tl
cor_data<- melt(cor_matrix)
```

```
# Visualize the Correlation Plot of the the Weather Parameters
ggplot(data=cor_data, aes(x= Var1, y= Var2, fill=value))+
  geom_tile(color= "white")+
  geom_text(aes(label = round(value, 2)), color = "black", size = 4) +
  labs(title="Weather Parameters, Age and Marathon Performance Correlation Pl
       x = "", y = "")+
   scale_fill_gradient2(low = "#075AFF",
                        mid = "#FFFFCC",
                        high = "#FF0000") +
  theme_bw()+
  theme(axis.text.x= element_text(angle =35, vjust= 1, hjust = 1),
    plot.title = element_text(hjust = 0.5))
# Scatterplot for fastest or best CR vs. age for men vs. women
new_labels <- c("0" = "Boston", "1" = "Chicago", "2" = "New York City", "3" =

cols <- c(Female = "#EE799F", Male = "#56B4E9")
ggplot(data = marathon, aes(y = actual_CR, x = Age, group = Sex)) +
  geom_point(aes(color = Sex), size = 0.05) +
  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 = "Association of course record with age and gender",
    subtitle = ""
  ) +
  theme(strip.background = element_rect(fill="lightblue", size=1, color="darl
        axis.title.y = element_text(size = 10),
        plot.title = element_text(size = 12, hjust = 0.5),
        legend.position = "bottom"
        )
# 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", "25-3
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 4: Co
       subtitle = "Comparison for course record distribution among age groups
  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),

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

```
ggplot(gender_combined) +
  geom_segment(data = Males, aes(x = WBGT_intervals, xend = Females$WBGT_inte
  geom_point(aes(WBGT_intervals, y = `Percent off course record`, color = Sex
 geom_text(aes(WBGT_intervals, y = `Percent off course record`, label = round
  theme_bw() +
  scale_color_manual(name = "Sex",breaks = c("Female", "Male"), values = cols
  ggtitle("Association between WBGT and marathon performance by gender") +
  labs(x = "WBGT", y = "Percent off course record (Median)") +
  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" = "#1874CD", "25-34" = "#B2DFEE", "35-44" = "#A5002

# Bin scatterplot
wbgt_dat2 <-  marathon %>%
  mutate(bin = ntile(`Wet Bulb Globe Temperature (WBGT)`, 35))

wbgt_dat2 <- wbgt_dat2 %>%
  group_by(bin, Age_group, Sex) %>%
  summarize(x_median = median(`Wet Bulb Globe Temperature (WBGT)`),
            y_median = median(`Percent off course record`), .groups = "drop"

wbgt_plot <- ggplot(wbgt_dat2, aes(x = x_median, y = y_median, color = Sex))
  geom_point(size = 0.5) +
  geom_smooth(aes(fill = Sex)) +
  scale_fill_manual(name = "Sex", values = cols) +
  scale_color_manual(name = "Sex", values = cols) +
  theme_bw() +
  facet_grid(~Age_group) +
  labs(x = "Wet Bulb Globe Temperature (WBGT)", y = "% 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),
        strip.background = element_rect(fill="lightblue", size=1, color="dark

# % relative humidity intervals
humid_dat <-  marathon %>%
```

```
  mutate(bin = ntile(`Percent relative humidity`, 35))

humid_dat <- humid_dat %>%
  group_by(bin, Age_group, Sex) %>%
  summarize(x_median = median(`Percent relative humidity`),
            y_median = median(`Percent off course record`), .groups = "drop")

humidity_plot <- ggplot(data = humid_dat, aes(x = x_median, y = y_median, col
  geom_point(size = 0.5) +
  geom_smooth(aes(fill = Sex)) +
  scale_fill_manual(name = "Sex", values = cols) +
  scale_color_manual(name = "Sex", values = cols) +
  theme_bw() +
  facet_grid(~Age_group) +
  labs(x = "% relative humidity", y = "% 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),
        strip.background = element_rect(fill="lightblue", size=1, color="dark

# Wind speed
Wind_dat <-  marathon %>%
  mutate(bin = ntile(`Wind speed`, 35))

Wind_dat <- Wind_dat %>%
  group_by(bin, Age_group, Sex) %>%
  summarize(x_median = median(`Wind speed`),
            y_median = median(`Percent off course record`), .groups = "drop")

Wind_plot <- ggplot(Wind_dat, aes(x = x_median, y = y_median, color = Sex)) +
  geom_point(size = 0.5) +
  facet_grid(~Age_group) +
  geom_smooth(aes(fill = Sex)) +
  scale_fill_manual(name = "Sex", values = cols) +
  scale_color_manual(name = "Sex", values = cols) +
  theme_bw() +
  labs(x = "Wind speed", y = "% 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),
```

```
        strip.background = element_rect(fill="lightblue", size=1, color="dark

# drybulb_plot
drybulb_dat <-  marathon %>%
  mutate(bin = ntile(`Dry bulb temperature`, 35))

drybulb_dat <- drybulb_dat %>%
  group_by(bin, Age_group, Sex) %>%
  summarize(x_median = median(`Dry bulb temperature`),
            y_median = median(`Percent off course record`), .groups = "drop")

drybulb_plot <- ggplot(drybulb_dat, aes(x = x_median, y = y_median, color = S
  geom_point(size = 0.5) +
  facet_grid(~Age_group) +
  geom_smooth(aes(fill = Sex)) +
  scale_fill_manual(name = "Sex", values = cols) +
  scale_color_manual(name = "Sex", values = cols) +
  theme_bw() +
  labs(x = "Dry bulb Temperature", y = "% 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),
        strip.background = element_rect(fill="lightblue", size=1, color="dark

# wetbulb_plot
wetbulb_dat <-  marathon %>%
  mutate(bin = ntile(`Wet bulb temperature`, 35))

wetbulb_dat <- wetbulb_dat %>%
  group_by(bin, Age_group, Sex) %>%
  summarize(x_median = median(`Wet bulb temperature`),
            y_median = median(`Percent off course record`), .groups = "drop")

wetbulb_plot <- ggplot(wetbulb_dat, aes(x = x_median, y = y_median, color = S
  geom_point(size = 0.5) +
  facet_grid(~Age_group) +
  geom_smooth(aes(fill = Sex)) +
  scale_fill_manual(name = "Sex", values = cols) +
  scale_color_manual(name = "Sex", values = cols) +
  theme_bw() +
  labs(x = "Wet bulb Temperature", y = "% 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),
        strip.background = element_rect(fill="lightblue", size=1, color="dark

# dew point
dp_dat <-  marathon %>%
  mutate(bin = ntile(`Dew point`, 35))

dp_dat <- dp_dat %>%
  group_by(bin, Age_group, Sex) %>%
  summarize(x_median = median(`Dew point`),
            y_median = median(`Percent off course record`), .groups = "drop")

dp_plot <- ggplot(dp_dat, aes(x = x_median, y = y_median, color = Sex)) +
  geom_point(size = 0.5) +
  facet_grid(~Age_group) +
  geom_smooth(aes(fill = Sex)) +
  scale_fill_manual(name = "Sex", values = cols) +
  scale_color_manual(name = "Sex", values = cols) +
  theme_bw() +
  labs(x = "Dew Point", y = "% 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),
        strip.background = element_rect(fill="lightblue", size=1, color="dark


# Solar radiation
solar_dat <-  marathon %>%
  mutate(bin = ntile(`Solar radiation in Watts`, 35))

solar_dat <- solar_dat %>%
  group_by(bin, Age_group, Sex) %>%
  summarize(x_median = median(`Solar radiation in Watts`),
            y_median = median(`Percent off course record`), .groups = "drop")

solar_plot <- ggplot(solar_dat, aes(x = x_median, y = y_median, color = Sex))
  geom_point(size = 0.5) +
  facet_grid(~Age_group) +
```

```
  geom_smooth(aes(fill = Sex)) +
  scale_fill_manual(name = "Sex", values = cols) +
  scale_color_manual(name = "Sex", values = cols) +
  theme_bw() +
  labs(x = "Solar Radiation in Watts", y = "% 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),
        strip.background = element_rect(fill="lightblue", size=1, color="dark


# Solar radiation
blackglobe_dat <-  marathon %>%
  mutate(bin = ntile(`Black globe temperature`, 35))

blackglobe_dat <- blackglobe_dat %>%
  group_by(bin, Age_group, Sex) %>%
  summarize(x_median = median(`Black globe temperature`),
            y_median = median(`Percent off course record`), .groups = "drop")

blackglobe_plot <- ggplot(blackglobe_dat, aes(x = x_median, y = y_median, col
  geom_point(size = 0.5) +
  facet_grid(~Age_group) +
  geom_smooth(aes(fill = Sex)) +
  scale_fill_manual(name = "Sex", values = cols) +
  scale_color_manual(name = "Sex", values = cols) +
  theme_bw() +
  labs(x = "Black Globe Temperature", y = "% 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 = 10),
        strip.background = element_rect(fill="lightblue", size=1, color="dark
ggarrange(wbgt_plot, drybulb_plot, wetbulb_plot, blackglobe_plot, nrow = 4, c
ggarrange(dp_plot,humidity_plot, Wind_plot, solar_plot, nrow = 4, common.lege
# Examine the association between environmental conditions and marathon perfo
# 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 = T
  select(Sex,`actual_CR`, `Dry bulb temperature`, `Wet bulb temperature`, `Pe
```

```
  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 marat
  theme(strip.background = element_rect(fill="lightblue", size=1, color="dark
        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
              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_si
```