Exploring the Impact of Environment and Weather Conditions on Marathon Performance

Tianna Chan

2024-10-06

Abstract

This project explores the impact of environmental factors—including temperature, humidity, solar radiation, wind, and air quality—on marathon performance across gender and age. Using data from major U.S. marathons (1993-2016), combined with weather and air quality information, the relationships between these factors and race times were examined. Results indicate significant age-related differences in air quality impacts. Additionally, wet bulb temperature and red flag conditions were associated to poorer performance, while solar radiation and dew point were associated with better performance. These findings provide valuable insights for athletes and coaches to optimize race-day strategies based on environmental conditions and individual characteristics.

Introduction

Previous studies have shown negative associations on endurance exercise performance and environmental temperature[1], which magnify during long distance events if temperature is warm [2]. It was also shown that the older population are less able to dissipate heat [3], and there are also sex differences [4].

This project aims to explore different environmental factors that impact marathon performance between gender and across age. This includes not only temperature, but humidity, solar radiation, wind, and air quality on a race date. Understanding the associations provides a more holistic comprehension on how environmental conditions influence performance in endurance events such as marathon. It could also be beneficial for athletes or coaches in forming their targeted strategies based on their own characteristics or weather.

The report starts with data preprocessing and showing some summary statistics for the data sets. To start the exploratory, the age and sex effects on marathon performance was first examined. Air quality information for PM2.5 and Ozone level was then factored in. Lastly,

weather parameters are investigated and a discussion on the results is included. All exploration was presented through plots, tables, and regression models.

The Data and Data Preprocessing

Project 1 Data

The `project1` data set contains info and results for 11564 participants (14-85 years) in the marathon races at Boston, New York City, Chicago, Twin Cities, and Grandmas from 1993 to 2016. It also includes 10 weather parameters, such as dry and wet bulb temperature, relative humidity, black globe temperature and more. This includes the calculated variables `WBGT` (Wet Bulb Globe Temperature, calculated by the three temperatures) and `Flag` (Groups for WBGT).

Using `str()`, the variable type for each column in the data set could be displayed. Column names included detailed information of the response, which was cleaned for readability and simplicity. Then, recoding was done to some columns to help ease the process of comprehending and merging later- The Race column was originally numerical with values 0, 1, 2, 3, and 4, and were recoded and factorized into B, C, NY, TC, D; The Sex column was originally numerical with values 0, and 1, recoded and factorized into "M" and "F". Additionally, Flag were factorized with White as the reference level. This is to make sure future analysis treat it the way we want. As histograms were plotted for all the numeric values in the data, we observe that only the %CR (percent off course record) was heavily right-skewed (See **Figure 1**).

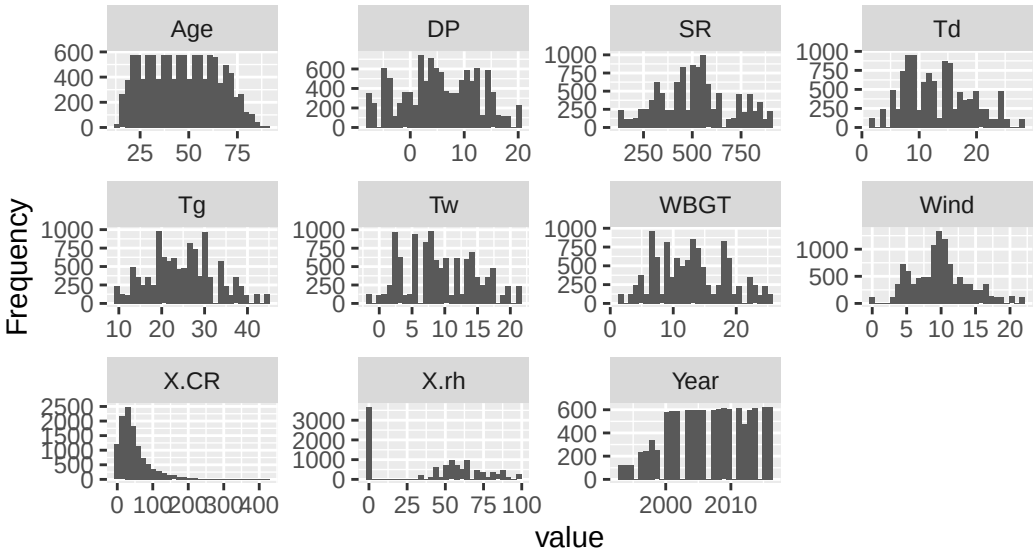

Figure 1: Histogram for Numeric Project 1 Data

To understand the data set, a missing data plot by `vis_miss()` was generated. The plot shows that there was a consistent 4% of missing data for the flag and weather measurement columns. This indicates that weather was not measured at these races. After some further exploration, table below shows that all data were missing at races 1 (Chicago), 2 (NYC), 3 (Twin Cities) at 2011, and race 4 (Grandma's) at 2012.

The flag for these were missing because they should have been calculated by the wet bulb globe temperature (WBGT), and the WBGT is also calculated by the other temperature variables, which were not measured and is dependent on these particular races in particular year (see **Table 1**). Therefore, the probability of the flag and WBGT being missing depends on both observed and unobserved variables, this should then be a case of **Missing Not at Random (MNAR).** For the other weather variables, they were missing because they are not measured in that particular year/race, but we do not have knowledge on the exact reason. Therefore, the probability of them being missing depends only on observed variables (year/race), they would then be a case of **Missing at Random (MAR).**

Table 1: Missing Data from Project 1 Data

Race	Year	Missing	Total
C	2011	126	126
D	2012	116	116
NY	2011	131	131
TC	2011	118	118

Air Quality Data

To further assess the environmental impact, AQI data was obtained using the R package RAQSAPI, grabbing data from the US Environmental Protection Agency's API. The code was provided in class and available at `aqi.R`. The resulted data set includes the arithmetic mean values for ozone level in parts per million, and the PM2.5 in Micrograms/cubic meter (LC).

Based on the same missing graph plotted, the AQI column has 16% missing values. Upon further analysis, it was found that all data with sample duration = 1 hour has the AQI column missing (See **Table 2**). This is then **Missing At Random (MAR)** since it only depends on this observed column.

Table 2: Missing Data from AQI Values

Sample Duration	Missing	Total
1 HOUR	1674	1674
24 HOUR	0	4034

Sample Duration	Missing	Total
24-HR BLK AVG	0	1335
8-HR RUN AVG BEGIN HOUR	0	3408

The summary table below (**Table 3**) uncover the complexity of this data set, showing the breakdown of information gathered on the data. Upon further review of the data, it was observed that the entries had different coding for marathon races, and it includes the full date for each race. Therefore, the marathon races were recoded to facilitate the merging of the data later, and the function `year()` was used to obtain the year only.

Table 3: AQI Data Summary

Parameter Code	Units of Measure	Sample Duration	n
44201	Parts per million	1 HOUR	1136
44201	Parts per million	8-HR RUN AVG BEGIN HOUR	3408
88101	Micrograms/cubic meter (LC)	1 HOUR	124
88101	Micrograms/cubic meter (LC)	24 HOUR	3964
88101	Micrograms/cubic meter (LC)	24-HR BLK AVG	930
88502	Micrograms/cubic meter (LC)	1 HOUR	414
88502	Micrograms/cubic meter (LC)	24 HOUR	70
88502	Micrograms/cubic meter (LC)	24-HR BLK AVG	405

According to the US Environmental Protection Agency, 88101 and 88502 are both used to report daily Air Quality Index values. The difference is that 88101 include both manually operated Federal Reference Methods (FRMs) and automated Federal Equivalent Methods (FEMs), but 88502 are "FRM-like" @EPA. Based on this information, only the 88101 data was taken for simplicity. PM2.5 below 12 g/m^3 is considered healthy with little to no risk from exposure, while anything above 35 g/m^3 is unhealthy @IndoorAirHygieneInstitute. For Ozone, it is measured under the parameter 44201. The quality standard is 0.08 ppm @EPA, anything above could be unhealthy. To clean the data, summary was created by taking the average of the arithmetic mean. However, note that there were incomplete data - it contains missing PM2.5 values from most of the early years and a few in latter years. It is possible that the PM2.5 values were not collected until later years. This could be a **Missing At Random (MAR)** case because it only depends on observed variable `year`.

The graph below on the left illustrates the change in ozone levels on race day, over the years, by marathon locations (See **Figure 2**). Each color/line represents a marathon location, allowing for a comparison of trends within those areas. The points represents the average ozone measurements, and different fluctuations between different locations was observed. The curve was plotted using `geom_smooth` that shows the trend. In particular, the average ozone levels

at New York City was the most stable throughout years; At Twin City we see the greatest decrease; The other locations have a slight fluctuations only. Using the same method, the graph on the right shows the trend on PM2.5. There were big fluctuations for all locations.

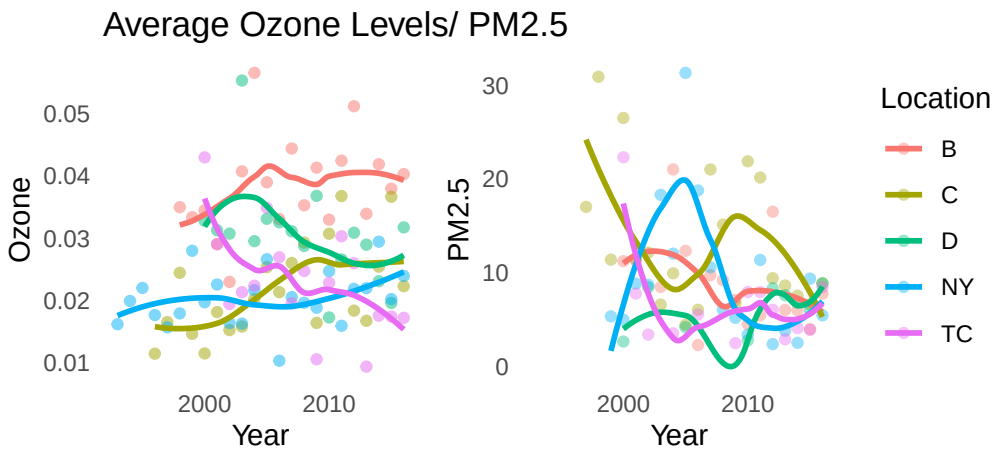

Figure 2: Average Ozone Levels/ PM2.5 by Race Location

Course Record Data

The project also utilized the `course_record` data that included the course record for each gender at each race and year. No missingness was found in the data.

Merging the Data

The course record data was merged to project 1 data using a `left_join()`. Afterwards, the record was then used to calculate a new variable `Time` for each participant, with the equation $\text{Time (hours)} = \frac{\text{CR} \times \left(1 + \frac{\%CR}{100}\right)}{3600}$ that adjusts for the percent off course record and convert the final unit to hours. Next, the AQI data was also joined using `left_join()`. **Figure 3** shows a correlation plot for all the numeric columns in the final data. High correlations were seen between some weather measures. It is also obvious that the course time was highly correlated with age. We will dive deeper and quantify some relations between the variables in the later sections.

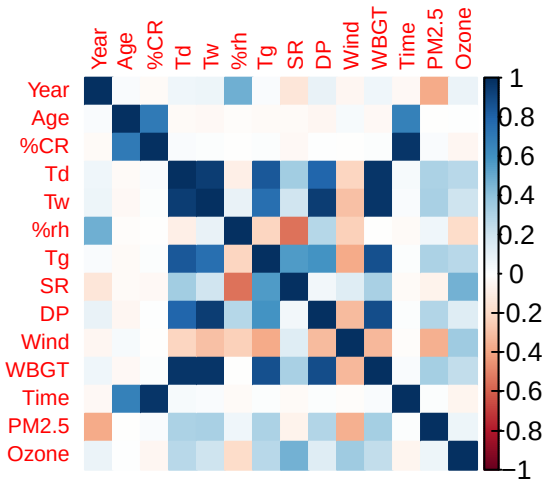

Figure 3: Correlation plot

Effects of increasing age on marathon performance in men and women

Figure 4 below shows a scatter plot of time and age for men and women. The plot is divided into two panels, by sex, and the blue line represent a fitted curve performed directly by `geom_smooth()`. There seem to be a near-identical trend for men and women- both sexes gets an increasingly faster time each year they are closer to around age 25. However, as they reach the fastest time at around age 25, they gradually run slower as age increases further. We can observe a spike happening around age 60 for both sexes, but men has a steeper slope. In other words, the slower in time was more significant in the male group. On the other hand, we can see the curve for women is slightly higher than that of the men. This means that women are running slightly slower in general.

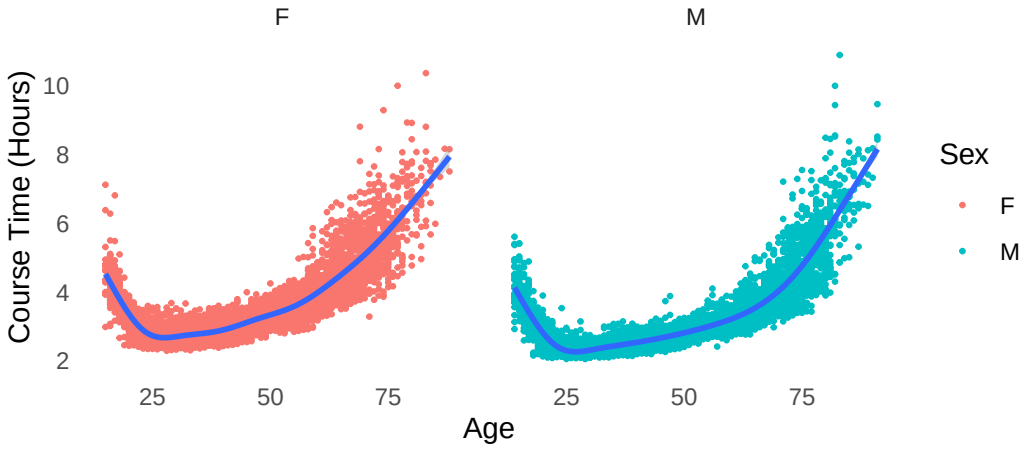

Figure 4: Course Time vs Age by Sex

This table shows the summary statistics for age and course time for men and women. It is obvious that men (N = 6,112) had a larger sample size than women (N = 5,452). In average, the men population is also older than the female population. We again see that men have a shorter course time on average.

Table 4: Age and Time Summary by Sex

Characteristic	**F**, N = 5,452	**M**, N = 6,112
Age	45 (30, 59)	48 (32, 64)
Time	3.55 (2.87, 3.91)	3.17 (2.49, 3.46)
[1] Mean (IQR)		

A regression model was fit to quantify the course time between sex and their ages.

$$E[\text{Time}] = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{SexM}$$

Based on the results below (rounded to 3 decimal places), both Age and Sex are significantly influential on time (p-value < 0.05), with age positively associated while sex negatively associated.

Table 5: Linear Regression Model Summary

Variable	**Estimate**	**Std. Error**	**t value**	**Significance**
(Intercept)	1.793	0.019	92.341	0
Age	0.039	0.000	104.527	0
SexM	-0.491	0.013	-36.523	0

Therefore, for each unit increase in Age, the expected Time increases by $\beta_1 = 0.039$ hours, holding all other variables constant. For **Male** (SexM = 1), the expected Time decreases by $\beta_2 = -0.491$ hours compared to females, holding all other variables constant.

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

The aggregated data was examined by calculating the average course time for each race/year (See **Figure 5**). Most races had PM2.5 on the lower end, about under 13. Just by looking at the graph, it does not immediately appear that there is a big difference in performance comparing different PM2.5 levels. The slope drawn by `geom_smooth()` slowly increase as average time increases. This suggests that as PM2.5 level goes up, the average course time goes up. In other words, they are running slower on average.

Same process for Ozone measurements, the slope plotted with `geom_smooth` shows a decreasing trend. This is strange because it suggests that lower ozone may be associated with a longer course time. In other words, runners tends to finish the race faster when the ozone level is higher. However, it is also noticeable from both graphs that the Boston Marathon has a lower average course time among all other marathon races, and their ozone measurements seems slightly above average.

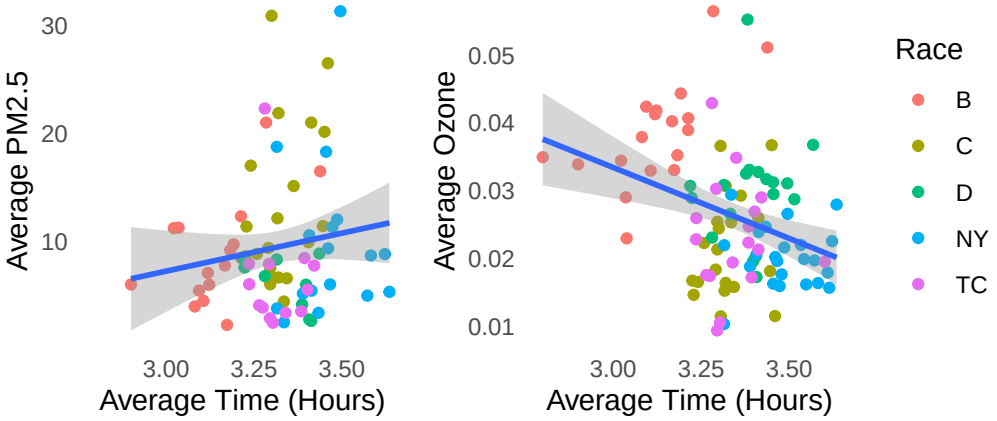

Figure 5: Average PM2.5/ Ozone vs. Average Course Time, by Race

In order to see the differences between sex, another aggregation by sex was created (See **Figure 6**). There was a clear distinction between male and female - again, male has a shorter average time than female in overall.

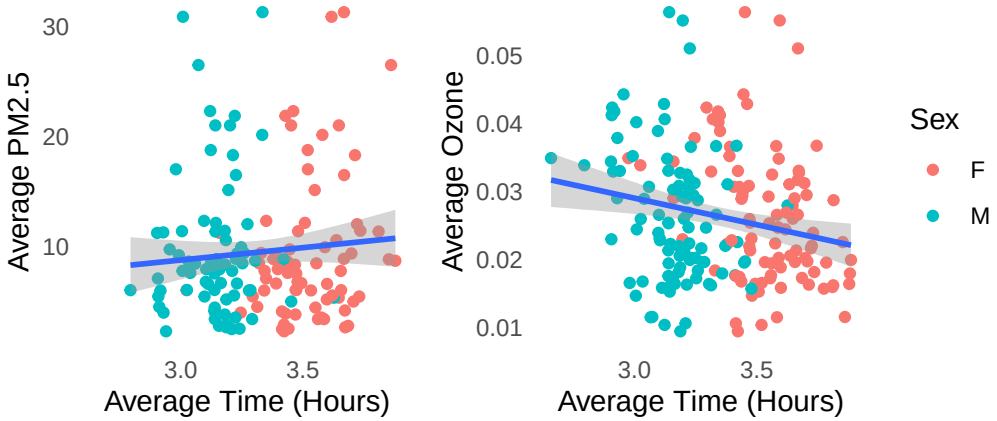

Figure 6: Average PM2.5/ Ozone vs. Average Course Time, by Sex

To quantify the impact of environmental conditions on marathon performance, let us first focus on a regression model with only PM2.5 and Ozone: $E[\text{Time}] = \beta_0 + \beta_1 \text{PM2.5} + \beta_2 \text{Ozone}$.

Table 6: Linear Regression Model Summary

Variable	Estimate	Std. Error	t value	Significance
(Intercept)	3.433	0.033	104.121	0.000
PM2.5	0.004	0.002	2.436	0.015
Ozone	-5.630	1.090	-5.167	0.000

Since both PM2.5 and Ozone are significant, both of them were able to explain course time.

$\beta_1 = 0.004$ means that for every one unit increase in PM2.5, the course time increases by 0.004 hours (14.4 seconds), holding others constant.

$\beta_2 = -5.630$ means that for every one unit increase in Ozone, the course time decreases by -5.630 hours. Thus, by conversion, for every 0.01 unit increase in Ozone, the course time decreases by -0.0563 hours (3.378 minutes).

To find out whether the impact differs across age and gender, the main effects of age and gender was added along with the two-way interactions of them with each environmental measure. Then, backward selection was used and the results are as follows. This method drops one variable each time and test for the significance.

The full model: $E[\text{Time}] = \beta_0 + \beta_1\text{Age} + \beta_2\text{SexM} + \beta_3\text{PM2.5} + \beta_4\text{Ozone} + \beta_5(\text{Age} \times \text{PM2.5}) + \beta_6(\text{Age} \times \text{Ozone}) + \beta_7(\text{SexM} \times \text{PM2.5}) + \beta_8(\text{SexM} \times \text{Ozone})$

Final model by backward selection: $E[\text{Time}] = \beta_0 + \beta_1\text{Age} + \beta_2\text{SexM} + \beta_3\text{PM2.5} + \beta_4\text{Ozone} + \beta_5(\text{Age} \times \text{PM2.5}) + \beta_6(\text{Age} \times \text{Ozone}) + \beta_7(\text{SexM} \times \text{PM2.5})$

Table 7: Linear Regression Model Summary

Variable	Estimate	Std. Error	t value	Significance
(Intercept)	1.7536	0.0651	26.9345	0.0000
Age	0.0409	0.0013	31.8361	0.0000
SexM	-0.4372	0.0264	-16.5871	0.0000
PM2.5	-0.0105	0.0034	-3.0985	0.0020
Ozone	4.1944	2.1464	1.9541	0.0507
Age:PM2.5	0.0004	0.0001	5.5247	0.0000
Age:Ozone	-0.2100	0.0428	-4.9104	0.0000
SexM:PM2.5	-0.0041	0.0023	-1.7602	0.0784

Since the interaction term for Sex * Ozone was removed, this suggests that the relationship between Ozone and Time does not differ significantly between males and females in the data set. The interaction term for Sex * PM2.5 was not removed by backward selection, but it remained insignificant (p = 0.078). We cannot conclude that the impact of PM2.5 on course

time changes as Sex group changes. The others were significant means that the impact of PM2.5 and ozone on course time changes as Age increases. In overall, the impact for these environmental factors were seemingly more significant to age differences, but not that much for different sexes. We also see an apparent reversal in the effects of PM2.5 and ozone, but they could be attributed to the inclusion of the interaction terms.

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

The graph below illustrates the relationship of WBGT, flag, and the marathon performance on course time. Seems like we can almost draw vertical lines to pinpoint the appearance of new flag colors at certain time points. To elaborate, only whites and one green had an average course time <3.12, then only whites and greens had an average course time <3.28, and only whites, greens, and yellows < 3.38. This suggests that larger WBGT may be associated with a longer course time.

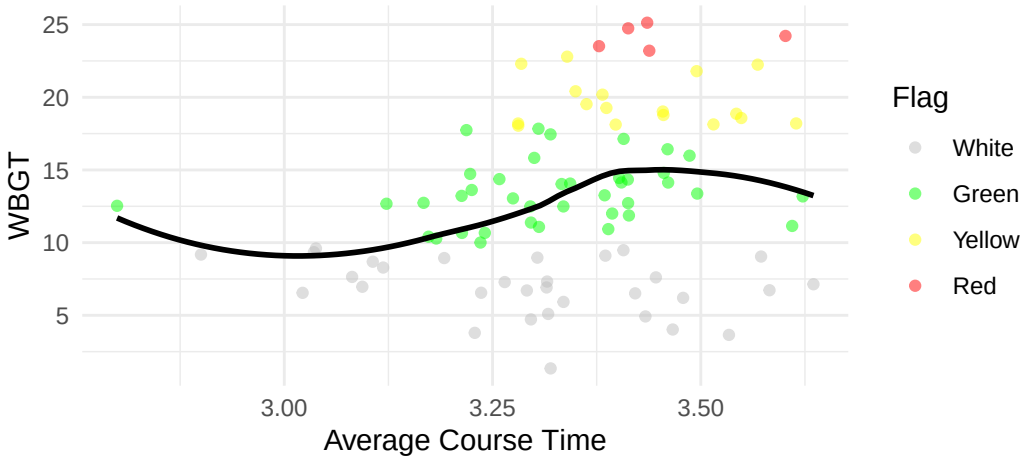

Figure 7: Impact of WBGT on Marathon Performance

Course time was regressed with the full model- all the participant characteristics and weather variables in the data set. Then, backward selection was used to find the smallest best model. **Table 8** below shows the coefficients of the final model by backward selection. This is the final model:

$$E[\text{Time}] = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 \cdot \text{SexM} + \beta_3 \cdot \text{FlagRed} + \beta_4 \cdot \text{FlagWhite} + \beta_5 \cdot \text{FlagYellow} + \beta_6 \cdot \text{Td} + \beta_7 \cdot \text{Tw} + \beta_8 \cdot \%\text{rh} + \beta_9 \cdot \text{SR} + \beta_{10} \cdot \text{DP}$$

Table 8: Linear Regression Model Summary: Backward Selection

Variable	**Estimate**	**Std. Error**	**t value**	**Significance**
(Intercept)	1.6982	0.0550	30.8900	0.0000
Age	0.0390	0.0004	102.8962	0.0000
SexM	-0.4919	0.0136	-36.1869	0.0000
FlagGreen	0.0281	0.0250	1.1243	0.2609
FlagYellow	0.1148	0.0436	2.6354	0.0084
FlagRed	0.1550	0.0620	2.5018	0.0124
Td	-0.0163	0.0092	-1.7803	0.0751
Tw	0.0589	0.0200	2.9386	0.0033
`%rh`	-0.0006	0.0003	-1.9177	0.0552
SR	-0.0002	0.0000	-4.2738	0.0000
DP	-0.0279	0.0090	-3.1050	0.0019

The significance for the flags verify what we have seen in the graph above. The Green Flag was not significant (p-value= 0.261), but the Yellow and Red were. This indicates that when the weather is under Green flag conditions, the runners' course time *do not* differ from those under White flag conditions. Under Yellow flag conditions, the runners have 0.115 hours (6.9 minutes) longer than those under White flag; and under Red flag conditions, the runners have 0.155 hours (9.3 minutes) longer than those under White flag.

The dry bulb temperature (Td) and humidity (%rh) were marginally significant, so their effect is still not strong enough to conclude anything.

The wet bulb temperature (Tw) is significant (p-value= 0.003), so the estimate 0.059 suggests that higher Tw lead to longer course time, which means worse performance.

Solar Radiation (SR) has an extremely small p-value, so the table rounded to 4 decimal places did not show the full value. The estimate -0.0002 indicates a significant negative association with average time that higher solar radiation leads to shorter time.

Lastly, Dew Point (DP) is significant (p-value= 0.0019), so the estimate of -0.028 indicates that higher dew points are associated with lower time, meaning better performance.

To summarize, **Wet Bulb Temperature (Tw) and Flag Conditions (especially Red)** have the largest positive impacts on time (i.e., largest negative impacts on performance). This means the hotter and more humid condition are associated with worse performance.

On the other hand, **Solar Radiation (SR) and Dew Point (DP)** has negative impact on time (i.e., positive impact on performance), so higher dew point could be a favorable condition for better performance.

Discussion

This project explored the impact of environmental and weather conditions such as PM2.5, ozone levels, temperature, and humidity on marathon performance across different age group and genders. The findings revealed significant associations between some of these variables and race times, which varied by age but less so by sex.

The data used in this analysis has a wide range of participants and races under different conditions. However, there are limitations that must be acknowledged. Specifically, the PM2.5 and Ozone levels were calculated by taking the average of all sample duration (e.g., 1-hour, 8-hour, and 24-hour measurements), which may not fully capture the variability of air quality experienced by participants throughout the marathon.

In conclusion, this analysis demonstrates that PM2.5, ozone levels, solar radiation, wet bulb temperature, red flag, and dew point were significantly associated with marathon performance. While ozone appeared to have an unexpected negative association with performance, it is possible that this reflects confounding factors there were not captured in this analysis.

References

1. Ely, B. R., Cheuvront, S. N., Kenefick, R. W., & Sawka, M. N. (2010). Aerobic performance is degraded, despite modest hyperthermia, in hot environments. Med Sci Sports Exerc, 42(1), 135-41.
2. Ely, M. R., Cheuvront, S. N., Roberts, W. O., & Montain, S. J. (2007). Impact of weather on marathon-running performance. Medicine and science in sports and exercise, 39(3), 487-493.
3. Kenney, W. L., & Munce, T. A. (2003). Invited review: aging and human temperature regulation. Journal of applied physiology, 95(6), 2598-2603.
4. Besson, T., Macchi, R., Rossi, J., Morio, C. Y., Kunimasa, Y., Nicol, C., … & Millet, G. Y. (2022). Sex differences in endurance running. Sports medicine, 52(6), 1235-1257.

Code Appendix

```
# Load libraries
library(readr)
library(tidyverse)
set.seed(123456)
library(knitr)
library(tidyr)
library(dplyr)
library(kableExtra)
library(visdat)
library(gtsummary)
library(patchwork)
library(DataExplorer)
# Import data sets
aqi_values <- read_csv("aqi_values.csv")
course_record <- read_csv("course_record.csv")
project1 <- read_csv("project1.csv")

###### DATA PREPROCESSING ######

## Project 1 Data

# Project 1 - Data Quality
# str(project1)

# Project 1 - Rename column names
colnames(project1) <- c("Race", "Year", "Sex", "Flag", "Age", "%CR", "Td",
 ↪  "Tw", "%rh", "Tg", "SR", "DP", "Wind", "WBGT")

# Project 1 - Recode columns
project1 <- project1 %>%
  mutate(Race = case_when(
    Race == 0 ~ "B",
    Race == 1 ~ "C",
    Race == 2 ~ "NY",
    Race == 3 ~ "TC",
    Race == 4 ~ "D",
    ),
    Sex = case_when(
    Sex == 0 ~ "F",
```

```
    Sex == 1 ~ "M"
    )
    ) %>%
  mutate(across(c(Race, Sex, Flag),
         as.factor))

# Project 1 - Set white flag as reference
project1$Flag <- factor(project1$Flag, levels = c("White", "Green", "Yellow",
 ↪  "Red"))

# FIGURE 1: Project 1 - Plot histogram
plot_histogram(project1, nrow = 5L)

# Project 1 - Missing data plot
# project1 %>% abbreviate_vars() %>% vis_miss()

# TABLE 1: Project 1 - Missing data table
project1 %>% group_by(Race, Year) %>%
  summarize(n_Miss = sum(is.na(Flag)),
            n_Total = n()) %>%
  filter(n_Miss > 0) %>%
  kable(digits = 2, caption = "Missing Data from Project 1 Data",
        col.names = c("**Race**", "**Year**", "**Missing**", "**Total**"))

## AQI Data
# AQI Data - Missing data plot
# vis_miss(aqi_values)

# TABLE 2: AQI Data - Missing data summary
aqi_values %>% select(c(sample_duration, aqi)) %>%
  group_by(sample_duration) %>%
  summarise(n_Miss = sum(is.na(aqi)),
            n_Total = n()) %>%
  kable(digits = 2, caption = "Missing Data from AQI Values",
        col.names = c("**Sample Duration**", "**Missing**", "**Total**"))

# TABLE 3: AQI Data - Data Quality
aqi_values %>%
  group_by(parameter_code, units_of_measure, sample_duration) %>%
  summarise(n=n()) %>%
  kable(caption = "AQI Data Summary",
        col.names = c("**Parameter Code**", "**Units of Measure**", "**Sample
         ↪  Duration**", "**n**"))
```

```r
# AQI Data - Change marathon names and add year column
aqi_values_adj <- aqi_values %>%
  mutate(
    marathon = case_when(
      marathon == "Boston" ~ "B",
      marathon == "Chicago" ~ "C",
      marathon == "NYC" ~ "NY",
      marathon == "Twin Cities" ~ "TC",
      marathon == "Grandmas" ~ "D"
      ),
    year = year(date_local)
    )

# AQI Data - Filter data for simplicity
aqi_values_adj <- aqi_values_adj %>%
  filter(parameter_code %in% c(88101, 44201)) %>%
  ↪  select(-c("cbsa_code","state_code","county_code","site_number","date_local"))

# AQI Data - Create summary by year, race
aqi_summary_by_year <- aqi_values_adj %>%
  group_by(year, marathon, units_of_measure) %>%
  summarise(avg_arithmetic_mean=mean(arithmetic_mean, na.rm = TRUE))

aqi_summary_by_year <-
  spread(aqi_summary_by_year, units_of_measure, avg_arithmetic_mean) %>%
  arrange(year)

colnames(aqi_summary_by_year) <- c("Year", "Race", "PM2.5", "Ozone")

# FIGURE 2: AQI - Change over time plots
a <- ggplot(aqi_summary_by_year, aes(x = Year, y = Ozone, color = Race)) +
  geom_point(alpha=0.5) +
  geom_smooth(se = FALSE) +
  # facet_wrap(~ Race) +
  labs(title = "Average Ozone Levels/ PM2.5",
       x = "Year") +
  theme_minimal() +
  scale_color_discrete(name = "Location") +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
```

```
  guides(color="none")
b <- ggplot(aqi_summary_by_year, aes(x = Year, y = `PM2.5`, color = Race)) +
  geom_point(alpha=0.4) +
  geom_smooth(se = FALSE) +
  # facet_wrap(~ Race) +
  labs(x = "Year") +
  theme_minimal() +
  scale_color_discrete(name = "Location") +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())

a + b

## Course Record
# Course Record - Missing data plot
# vis_miss(course_record)

# Course Record - Data Quality
course_record <- course_record %>% rename("Sex" = "Gender")
## Merging the Data
# Merge project 1 with course record
project1_CR <- project1 %>%
  left_join(course_record, by = c("Race", "Year", "Sex"))

# Calculate time
project1_CR$Time <- (project1_CR$CR * (1+project1_CR$`%CR`/100))/3600 %>%
  as.numeric()

# Merge project 1 + course record with AQI data
project1_CR_aqi <- project1_CR %>%
  left_join(aqi_summary_by_year, by = c("Race", "Year"))

# FIGURE 3: Correlation of full numeric data
library(corrplot)

project1_CR_aqi$Time <- project1_CR_aqi$Time %>% as.numeric()
data <- select_if(project1_CR_aqi,is.numeric) %>% na.omit() %>%
 ↪  abbreviate_vars()
M = cor(data)
corrplot(M, method = 'color', tl.cex = 0.7)

###### AIM 1 ######
```

```r
# FIGURE 4: Age vs Time plot
ggplot(project1_CR_aqi, aes(x = Age, y = Time)) +
  geom_point(aes(color=Sex), size = 0.5) +
  geom_smooth() +
  facet_wrap(~ Sex) +
  labs(x = "Age",
       y = "Course Time (Hours)") +
  theme_minimal() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())

# TABLE 4: Age & Time Summary by Sex
project1_CR_aqi %>% dplyr::select(Age, Sex, Time)%>%
  tbl_summary(by = Sex,
              type = list(where(is.numeric) ~ "continuous"),
              statistic = list(all_continuous() ~ "{mean} ({p25}, {p75})"))
               ↪  %>%
  as_kable_extra(booktabs = TRUE,
                 caption = "Age and Time Summary by Sex",
                 longtable = TRUE, linesep = "") %>%
  kableExtra::kable_styling(font_size = 11,
                            latex_options = c("repeat_header",
                             ↪  "HOLD_position"))

# Regression model - Time effects on Sex and Age
model <- lm(as.numeric(Time) ~ Age + Sex, data = project1_CR_aqi)

# TABLE 5: Print results
summary <- summary(model)
coefficients <- as.data.frame(summary$coefficients)
kable(coefficients, digits = 3,
      caption = "Linear Regression Model Summary",
      col.names = c("**Variable**", "**Estimate**", "**Std. Error**", "**t
       ↪  value**", "**Significance**"))

###### AIM 2 ######

# Create summary data by race & year
new_data <- project1_CR_aqi %>% group_by(Race, Year, PM2.5, Ozone) %>%
 ↪  summarize(Avg_Time = mean(as.numeric(Time)))
```

```r
# FIGURE 5: AQI data vs Average Time plots
a <- ggplot(new_data, aes(x = Avg_Time, y = PM2.5)) +
  geom_point(aes(color = Race)) +
  geom_smooth(method = "lm") +
  labs(x = "Average Time (Hours)",
       y = "Average PM2.5") +
  theme_minimal() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  guides(color="none")

b <- ggplot(new_data, aes(x = Avg_Time, y = Ozone)) +
  geom_point(aes(color = Race)) +
  geom_smooth(method = "lm") +
  labs(x = "Average Time (Hours)",
       y = "Average Ozone") +
  theme_minimal() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())

a + b

# Create summary table further by sex
new_data <- project1_CR_aqi %>% group_by(Race, Year, Sex, PM2.5, Ozone) %>%
 ↪  summarize(Avg_Time = mean(as.numeric(Time)))

# FIGURE 6: AQI data vs Average Time plots BY SEX
a <- ggplot(new_data, aes(x = Avg_Time, y = PM2.5)) +
  geom_point(aes(color = Sex)) +
  geom_smooth(method = "lm") +
  labs(x = "Average Time (Hours)",
       y = "Average PM2.5") +
  theme_minimal() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  guides(color="none")

b <- ggplot(new_data, aes(x = Avg_Time, y = Ozone)) +
  geom_point(aes(color = Sex)) +
  geom_smooth(method = "lm") +
  labs(x = "Average Time (Hours)",
       y = "Average Ozone") +
```

```
  theme_minimal() +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank())

a + b

# Regression model, Time to AQI data
model <- lm(Time ~ PM2.5 + Ozone, data = project1_CR_aqi)

# TABLE 6: Print results
summary <- summary(model)
coefficients <- as.data.frame(summary$coefficients)
kable(coefficients, digits = 3,
      caption = "Linear Regression Model Summary",
      col.names = c("**Variable**", "**Estimate**", "**Std. Error**", "**t
       ↪  value**", "**Significance**"))

# Regression model, Time to AQI data + Age & Sex
full_model <- lm(Time ~ Age + Sex + PM2.5 + Ozone
                 + Age * PM2.5 + Age * Ozone
                 + Sex * PM2.5 + Sex * Ozone,
                 data = project1_CR_aqi)

# Backward selection
backward_model <- step(full_model, direction = "backward")
summary <- summary(backward_model)

# TABLE 7: Print results
coefficients <- as.data.frame(summary$coefficients)
kable(coefficients, digits = 4,
      caption = "Linear Regression Model Summary",
      col.names = c("**Variable**", "**Estimate**", "**Std. Error**", "**t
       ↪  value**", "**Significance**"))

###### AIM 3 ######

# Create summary table by race, year, WBGT, flag
new_data <- project1_CR_aqi %>% group_by(Race, Year, WBGT, Flag) %>%
 ↪  summarize(Avg_Time = mean(as.numeric(Time))) %>% na.omit()

# FIGURE 7: Plot average time by WBGT/Flag
ggplot(new_data, aes(x = Avg_Time, y = WBGT)) +
```

```
  geom_point(aes(color = Flag), alpha = 0.5) +
  geom_smooth(method = "loess", se = FALSE, color = "black") +
  labs(x = "Average Course Time",
       y = "WBGT",
       color = "Flag") +
  scale_color_manual(values = c("Green" = "green",
                                "Red" = "red",
                                "White" = "grey",
                                "Yellow" = "yellow"
                               )) +
  theme(panel.grid.major = element_blank(),
        panel.grid.minor = element_blank()) +
  theme_minimal()

# Regression model - Course Time & Weather
model <- lm(as.numeric(Time) ~ Age + Sex + Flag + Td + Tw + `%rh` + Tg + SR +
 ↪  DP + Wind + WBGT, data = project1_CR_aqi)

# Backward selection model
final_model <- step(model, direction = "backward")
summary <- summary(final_model)

# TABLE 8: Print results
coefficients <- as.data.frame(summary$coefficients)

kable(coefficients, digits = 4,
      caption = "Linear Regression Model Summary: Backward Selection",
      col.names = c("**Variable**", "**Estimate**", "**Std. Error**", "**t
       ↪  value**", "**Significance**"))
```